(2)

George Mason University

# EVALUATION AND ESTIMATION OF HANDLING QUALITIES VIA STATISTICAL MODELING OF PILOT RESPONSE DATA

Donald T. Gantz
Lawrence C. Baldwin
Linda J. Davis

Center for Computational Statistics
George Mason University
Fairfax, Va 22030

November, 1991

FINAL REPORT

Prepared for:
Office of Naval Research
800 N. Quincy Street
Arlington, VA 22217-5000

ONR Grant N00014-89-J-3146

# OVERVIEW

This report describes a research project which measured pilot response to seven control systems simulating different handling qualities, quantitatively evaluated and compared the systems based on these measurements, and compared the quantitative system evaluation based on measured pilot performance with a qualitative evaluation using the Cooper-Harper technique.

The objective of the project is implementation of a methodology for system evaluation via pilot performance to complement the current evaluation technique based on subjective ratings by test pilots. Pilot performance is determined through analysis of objective dynamic measurements of pilot response typical of flight test environments. In short, the methodology specifies a general approach fcr condensing the typically huge mound of measured test data accumulated during flight simulation experiments into meaningful quantities for system evaluation. The key element in the methodology is statistical modeling of a law for pilot control. Statistical modeling of pilot control provides an assessment of pilot performance in terms of standard statistical estimation parameters. The methodology requires that this control model be used to compute control input in a closed loop tracking task; the accuracy of the control model in performing this task is an important measure of pilot performance relevant to system evaluation. In addition, these parameters computed from the dynamic measurements of pilot performance are shown to enhance understanding of the aspects of the handling qualities underlying subjective rating techniques such as Cooper-Harper.

# TABLE OF CONTENTS

SECTION I INTRODUCTION . . . 1

SECTION II DESCRIPTION OF THE EXPERIMENT . . . 4
A. Experimental Setup . . . 4
B. Procedure . . . 9
C. Experimental Measurements . . . 9
D. Lead and Lag . . . 10

SECTION III PILOT RESPONSE MEASUREMENTS . . . 14
A. Overview . . . 14
B. Control Law Modeling . . . 15
C. Primary Measures of Pilot Performance . . . 27
D. Secondary Measures of Pilot Performance . . . 31
E. Scaling Associated with Screen Vertical Display Scale and Bandwidth of $\theta_c$ . . . 35
F. Summary of Pilot Performance Measures . . . 36

SECTION IV RELATIONSHIP BETWEEN PERFORMANCE MEASURES AND SYSTEMS . . . 37
A. Overview . . . 37
B. General Characteristics of the Data . . . 37
B.1. Primary Performance Measures . . . 38
B.2. Secondary Performance Measures . . . 43
C. Joint Analysis of Performance Measures using Discrimination Analysis . . 47
C.1. Discrimination Algorithms . . . 47
C.2. Application of Statistical Discrimination Algorithms - STEPDISC and CANDISC . . . 48
C.3. Application of Statistical Discrimination Algorithms - CART . . . . 54
D. Evaluation of Handling Qualities through the Analysis of Performance Measures . . . 64

SECTION V RELATIONSHIP BETWEEN COOPER-HARPER RATING AND PERFORMANCE MEASURES . . . 73
A. Subjective Pilot Opinion - Cooper-Harper Rating . . . 73
B. Correlation Between Cooper-Harper Rating and Performance Measures . . . 74
C. Evaluation of Handling Qualities . . . 77

SECTION VI CONCLUSIONS . . . 79

REFERENCES . . . 80

## TABLE OF CONTENTS (Continued)

APPENDIX A PILOT FLIGHT EXPERIENCE QUESTIONNAIRE . . . . . . . . . . . . . . 81

APPENDIX B RUN SEQUENCES EXECUTED BY PILOTS . . . . . . . . . . . . . . . . 83

APPENDIX C SUBJECTIVE EVALUATION PROCEDURE USING COOPER-HARPER RATING . . . . . . . . . . . . . . . . . . . . . . . . 92

APPENDIX D STATISTICAL DISCRIMINATION ALGORITHMS . . . . . . . . . . . . . . 94
D-1. SAS Procedure STEPDISC . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 94
D-2. SAS Procedure CANDISC . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 94
D-3. CART (Classification and Regression Trees) . . . . . . . . . . . . . . . . . . . 95

## LIST OF FIGURES

Figure 1. Relationships between Pilot Response, Pilot Opinion and Handling Qualities . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3
Figure 2. Primary Tracking Task Display . . . . . . . . . . . . . . . . . . . . . . . 4
Figure 3. Primary Tracking Task with Secondary Loading Task . . . . . . . . . . . . . 7
Figure 4. Scatter Plots of $\delta_e$ against $\theta_e$, with $\theta_e$ Lagged by 0.40 Seconds . . . . . . . 12
Figure 5. Linear Regression Fits and Residual Plots ($\delta_e$ against Lagged $\theta_e$) . . . . . 16
Figure 6. A Saturated Control Law . . . . . . . . . . . . . . . . . . . . . . . . . . 15
Figure 7. PROC NLIN Output . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 21
Figure 8. Nonlinear (Logistic Curve) Regression Fits and Residual Plots ($\delta_e$ against Lagged $\theta_e$) . . . . . . . . . . . . . . . . . . . . . . . . . . 23
Figure 9. Calculation of Performance Measures Describing Errors in Maintaining a Desired Control Law . . . . . . . . . . . . . . . . . . . . . . . . . 28
Figure 10. Calculation of the Performance Measure Describing How Accurately the Desired Control Law Performs the Primary Tracking Task . . . . . . 29
Figure 11. Tracking of the Input Signal by the Estimated Control Law . . . . . . . . . 30
Figure 12. Examples of the Variability of Instability Parameter $\lambda_s$ . . . . . . . . . . . . 32
Figure 13. Physical Characteristics of $P_1$, $P_2$, $P_3$ and $P_4$ . . . . . . . . . . . . . . . . . 33
Figure 14. Logistic Curves for Various Values of $P_1$, $P_2$, $P_3$ and $P_4$ . . . . . . . . . . . 34
Figure 15. Box Plot Example for a Set of Data . . . . . . . . . . . . . . . . . . . . . . 38
Figure 16. Regression Mean Square (Scaled); Box Plot Summary by Systems . . 39
Figure 17. Error Mean Square (Scaled); Box Plot Summary by System . . . . . . . . 40
Figure 18. R-Square; Box Plot Summary by Systems . . . . . . . . . . . . . . . . . 41
Figure 19. Mean Absolute Tracking Error From Data; Box Plot Summary by Systems . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 41
Figure 20. Mean Absolute Tracking Error From Simulation; Box Plot Summary by Systems . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 42
Figure 21. Maximum of $\lambda_s$; Box Plot Summary by Systems . . . . . . . . . . . . . . . 43
Figure 22. Mean of $\lambda_s$; Box Plot Summary by Systems . . . . . . . . . . . . . . . . . 44
Figure 23. Control Law Parameter $P_1$; Box Plot Summary by Systems . . . . . . . . 44
Figure 24. Control Law Parameter $P_2$; Box Plot Summary by Systems . . . . . . . . 45
Figure 25. Control Law Parameter $P_3$; Box Plot Summary by Systems . . . . . . . . 46
Figure 26. Control Law Parameter $P_4$; Box Plot Summary by Systems . . . . . . . . 46
Figure 27. Classification Tree Diagram; Tree with Seven Terminal Nodes . . . . . . 55
Figure 28. Splitting Rule Information from CART Output . . . . . . . . . . . . . . . 56
Figure 29. Geometric View of CART Terminal Node Classes . . . . . . . . . . . . . . 58
Figure 30. Estimated Control Laws for Correctly Classified Runs . . . . . . . . . . . 67
Figure 31. Median Control Laws for Each System (Excluding $K/s^2$) . . . . . . . . . . . 70
Figure 32. Control Law Parameter $P_3$; Box Plot Summary by System and CART Terminal Node; Correctly Classified Pilot Runs Only . . . . . . . . . . . . 69
Figure 33. Control Law Parameter $P_3$; Box Plot Summary by System and CART Terminal Node; Misclassified Groups Indicated by '*' . . . . . . . . . . . 71

## LIST OF FIGURES (Continued)

Figure 34. Cooper-Harper Rating; Box Plot Summary by Systems . . . . . . . . . . . . 73
Figure 35. Cooper-Harper Rating; Box Plot Summary by System and CART Terminal Node; Misclassified Groups Indicated by '*' . . . . . . . . . . . . 77

## LIST OF TABLES

Table 1. Tracking Target Pitch Frequencies and Phases . . . . . 5
Table 2. System and Gain Configurations . . . . . 8
Table 3. Variables Sampled and Recorded During Each Run for Each Pilot . . . . 10
Table 4. Discrimination Power of Performance Measures Measured by Squared Correlation and Squared Partial Correlation with Systems . . 49
Table 5. Summary of Canonical Discrimination Analysis; Various Sets of Performance Measures . . . . . 51
Table 6. Significantly Different Systems Based on Mahalanobis Distance Using Various Performance Measure Sets . . . . . 52
Table 7. System Medians for the Canonical Variables . . . . . 53
Table 8. Classification Matrix for Pilot Runs . . . . . 60
Table 9. Cross-Validation Classification Matrix for Pilot Runs . . . . . 60
Table 10. Breakdown of Pilot Runs by CART Terminal Node and Design Parameters for SYSTEM=K/s(s+1) . . . . . 62
Table 11. Breakdown of Pilot Runs by CART Terminal Node and Design Parameters for SYSTEM=K/s(s+2) . . . . . 62
Table 12. Breakdown of Pilot Runs by CART Terminal Node and Design Parameters for SYSTEM=K/s(s+4) . . . . . 62
Table 13. Breakdown of Pilot Runs by CART Terminal Node and Design Parameters for SYSTEM=K/s . . . . . 63
Table 14. Breakdown of Pilot Runs by CART Terminal Node and Design Parameters for SYSTEM=Poly1 . . . . . 63
Table 15. Breakdown of Pilot Runs by CART Terminal Node and Design Parameters for SYSTEM=Poly2 . . . . . 64
Table 16. Breakdown of Pilot Runs by CART Terminal Node and Design Parameters for SYSTEM=K/s$^2$ . . . . . 64
Table 17 Correlation Between Cooper-Harper Rating and Performance Measures . . . . . 74
Table 18. Discrimination Power of Performance Measures Measured by Squared Correlation and Squared Partial Correlation with Cooper-Harper Rating (Rounded) . . . . . 75
Table 19. Medians of the Canonical Variable per Cooper-Harper Group . . . . . 76
Table A-1. Experience Summary . . . . . 82
Table C-1. Adapted Cooper-Harper Rating Scale . . . . . 93

# SECTION I
# INTRODUCTION

Manual control of an aircraft involves a series of judgments by the pilot in order to perform some specified sequence of maneuvers. The pilotage task requires the pilot to extract information from a display, evaluate and estimate the error state of the aircraft with respect to the assigned task, and respond with control input intended to minimize the error. Handling qualities, as applied to piloted aircraft, is the preferred term to describe the characteristics of the aircraft that govern its controllability by the pilot. Controllability is the ease and precision by which the pilot is able to perform the tasks of aircraft control. For this research project, seven simulated flight systems were designed to demonstrate different handling qualities in a simple tracking task. Test pilots flew these simulated systems under experimental conditions. These pilots rated the systems on a subjective scale, and data from the experiments were objectively analyzed.

Historically, a subjective assessment by the pilot has been the principal method for evaluation of handling qualities. Numerical rating scales have been devised for the purpose of quantifying the subjective assessment by test pilots of their perceived performance. Over many years, the use of these subjective rating scales has been refined both in theory and practice. Their shortcomings have been in two areas. First, replication of subjective ratings is a persistent problem. Second, subjective rating scales do not provide objective information to engineers as to defects and attributes of their designed systems.

The objective of this research project is a demonstration of objective evaluation of handling qualities using recorded pilot response. In flight simulations as well as in actual flight tests, computers record detailed dynamic information and pilot response. The research reported here concerns the statistical analysis of pilot response to develop performance measures for the evaluation of handling qualities. The statistical analysis of the dynamic information centers around modeling the pilot's control law in the sense that a "mathematical" law for pilot control input is estimated from the test data. The model control law is then used to "fly" the same system, and the performance of the model is then measured. Statistical quantities (called parameters) of the modeling process and error measures for the model itself make up the primary performance measures defined in this research.

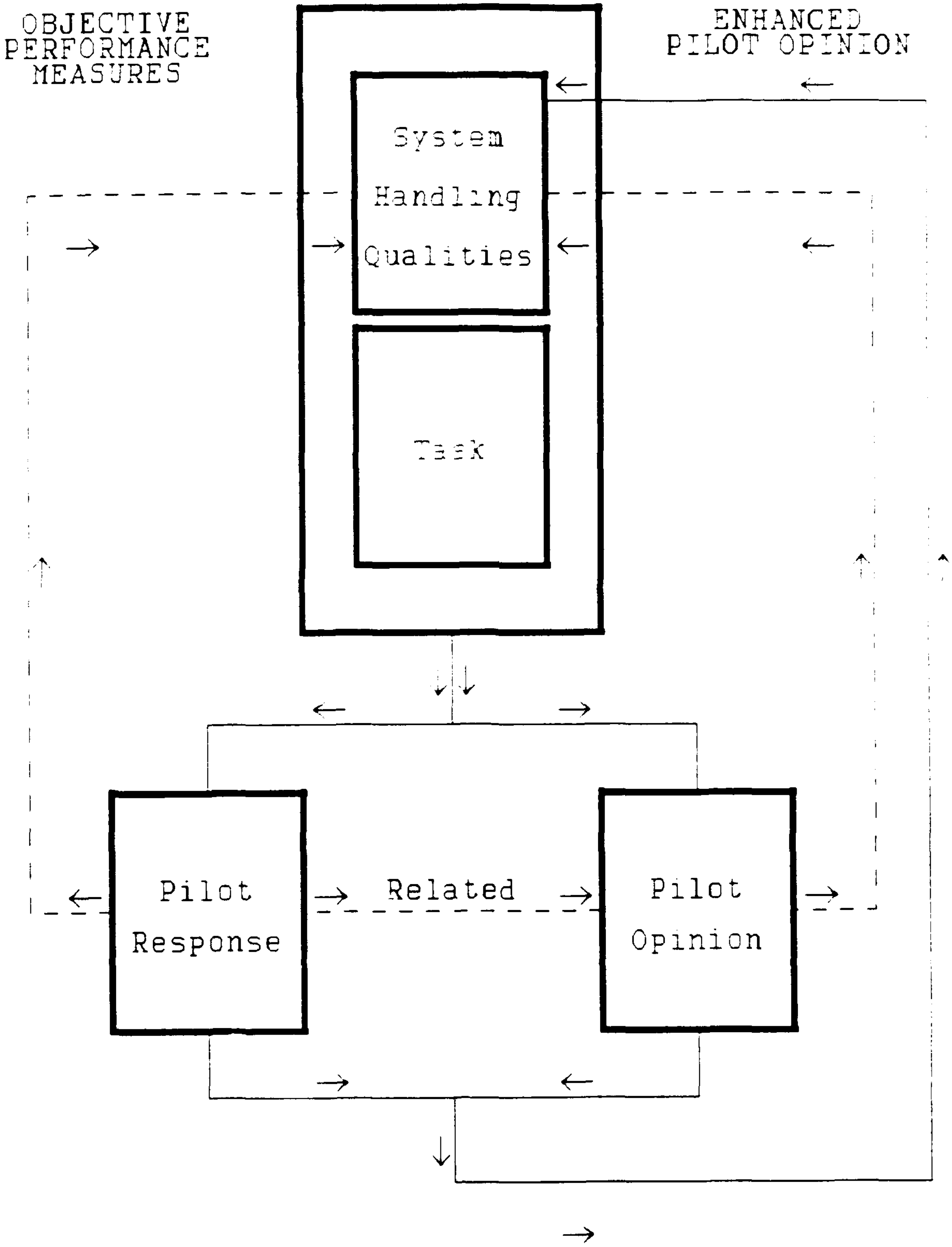
These performance measures are shown to effectively differentiate between the seven physical systems evaluated. The performance measures provide objective information regarding both pilot control techniques and effectiveness in performing the tracking task. This demonstration is an application of a methodology developed by the authors in their earlier work; see Baldwin and Gantz (1983, 1984).

Additionally, the performance measures from the statistical analysis of pilot response are shown to be related to the subjective pilot opinion ratings. This relationship between objective analysis of pilot response and subjective pilot opinion is exploited to enhance the value of pilot opinion for the evaluation of handling qualities. The diagram in Figure 1 illustrates the relationships between pilot response, pilot opinion, and the evaluation of handling qualities.

Thus, the objective performance measures based on statistical analysis of pilot response prove to be an informative and useful complement to subjective performance measures. The objective performance measures capture the characteristics of pilot response while performing the tracking task. The captured characteristics are those which influence a pilot's subjective evaluation of the controlled system as shown through the correlation of these objective measures with the subjective Cooper-Harper rating.

The finer details of the control law modeling and the statistical analysis in this project are sometimes artful and specifically dependent on the systems defined for this research. However, the methodology including the generic definitions of the major performance measures is broadly applicable.

Figure 1. Relationships between Pilot Response, Pilot Opinion and Handling Qualities

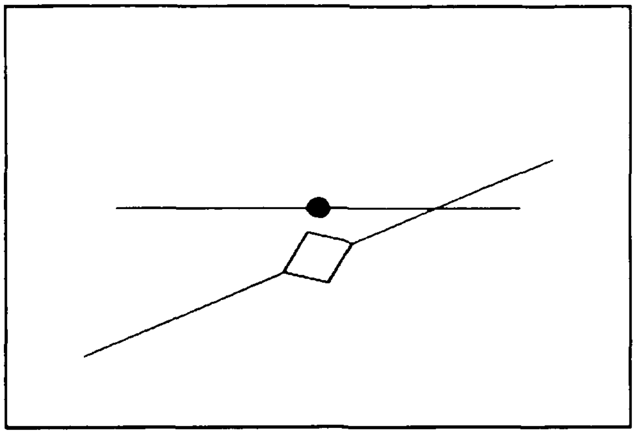


# SECTION II
# DESCRIPTION OF THE EXPERIMENT

## A. Experimental Setup

The simulation is programmed on a Silicon Graphics IRIS 3000 computer to provide display and tracking dynamics. An attitude bar symbol (colored white) displays pitch and roll relative to an horizon in the same sense as an aircraft artificial horizon. The tracking target is a red bar-circle symbol representing an aircraft ahead of and in the view of the pilot. A sketch of the display is shown in Figure 2. The display geometry is similar to an aircraft gun sight arrangement. Two vertical display scales were used: "normal" and "wide". The "normal" vertical display scale is a three centimeter displacement of the target representing about four degrees at the pilot's eye. The "wide" vertical display scale is approximately double the "normal" scale. The pilot control device system uses a commercial two-axis hand controller. The control grip is a spring-restrained, center-located unit. The hand controller is conveniently located near the pilot's right hand. Spring gradients are light; however, the unit "feel" is similar to a controller installed in a military simulator.

**Figure 2.** Primary Tracking Task Display



The tracking task used is similar to that in McDonnell (1968). It consists of a primary task of tracking a target as accurately as possible while performing a secondary task of maintaining wings level. The pitch of the tracking target is computed as a sum of twelve sinusoids, modulated by phase shifts in some of the lower frequency components:

$$\theta_c(t) = A_1 \sum_{i=1}^{n} \sin(\omega_i t + \zeta_i) + A_2 \sum_{i=n+1}^{12} \sin(\omega_i t + \zeta_i) \qquad (1)$$

where

$\theta_c$ = Pitch of the Tracking Target
$\omega$ = Angular Frequency
$\zeta$ = Phase Angle
$n$ = Bandwidth
$A_1$ = 0.10
$A_2$ = 0.02

Phase shifts are inserted in order to compensate for some unrealistically rapid changes in target pitch. The frequencies and phase shifts used are listed in Table 1; the number of cycles per 100 seconds is also given.

Two bandwidths are used. The "low" bandwidth, 1.885 radians per second, contains components 1 through 6 (i.e., n=6 in Equation 1) with "shelf" containing components 7 through 12. The "high" bandwidth, 4.775 radians per second, contains components 1 through 8 (i.e., n=8 in Equation 1) with "shelf" containing 9 through 12. The sinusoids making up the "shelf" have amplitude 14 db down from the amplitude of the lower frequencies within the bandwidth.

**Table 1.** Tracking Target Pitch Frequencies and Phases

| Component Number | Angular Frequency $\omega$ (radians/sec) | Cycles per 100 sec $100\omega/2\pi$ | Phase $\zeta$ (radians) |
|---|---|---|---|
| 1 | 0.188 | 3 | 0 |
| 2 | 0.251 | 4 | $\pi$ |
| 3 | 0.565 | 9 | $\pi/2$ |
| 4 | 0.754 | 12 | $\pi$ |
| 5 | 1.131 | 18 | $\pi/2$ |
| 6 | 1.885 | 30 | $\pi$ |
| 7 | 2.890 | 46 | 0 |
| 8 | 4.775 | 76 | 0 |
| 9 | 7.351 | 117 | 0 |
| 10 | 9.236 | 147 | 0 |
| 11 | 12.252 | 195 | 0 |
| 12 | 15.017 | 239 | 0 |

The primary tracking task is displayed to the pilot on the computer monitor as shown in Figure 2.

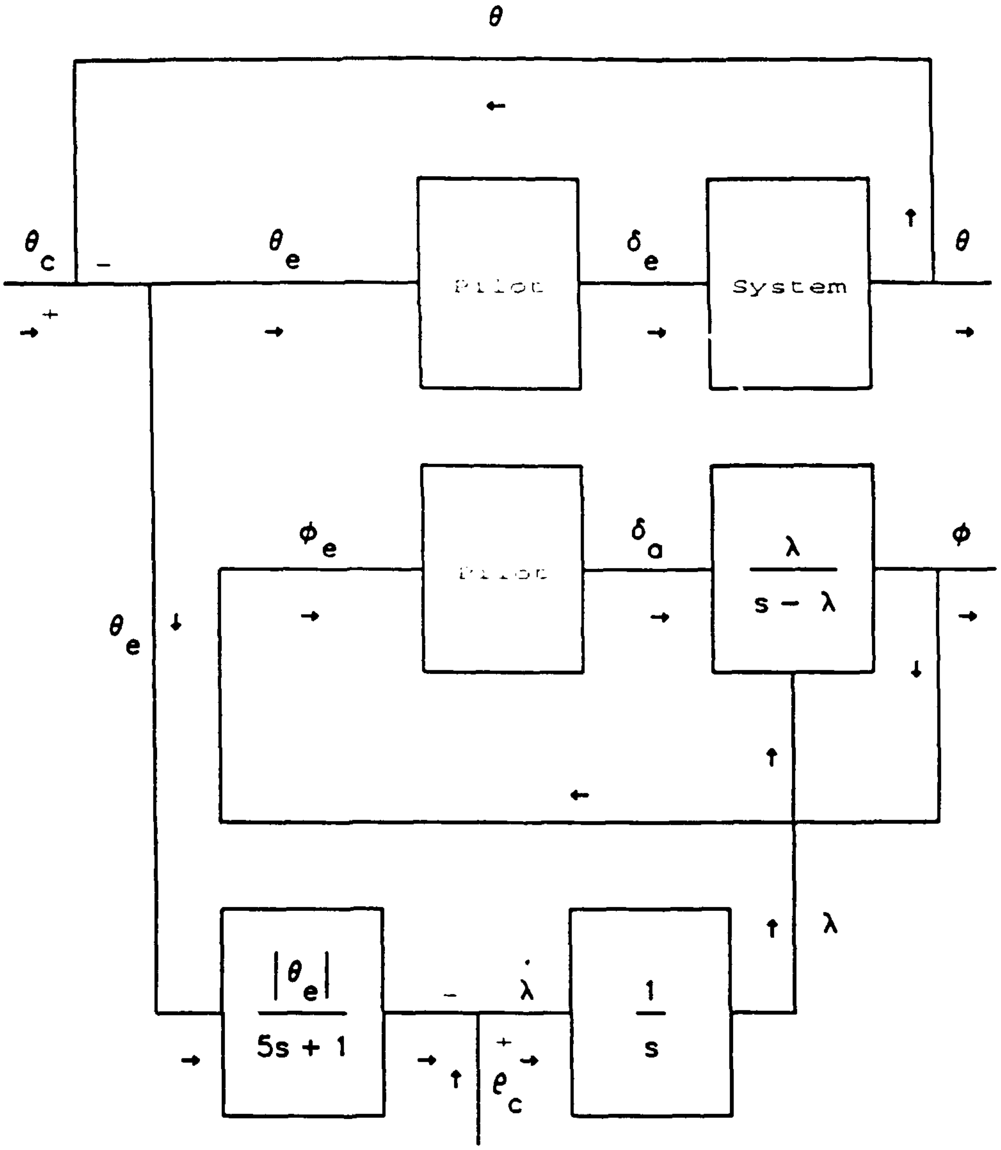
A secondary loading :ask is introduced in order to force the pilot to reach his capacity in performing the primary tracking task. The secondary task is an unstable tracking task in roll based on the "critical" task developed by Jex, McDonnell, and Phatak (1966). The task requires the pilot to maintain "wings level" ($\phi=0$), while performing the primary tracking task. Both the primary tracking task and the secondary task follow the scheme used by McDonnell (1968) and are shown in Figure 3. The roll loop develops an unstable root related to the primary tracking task difficulty and to the pilot's time delay.

To assure that the tracking of the target remains the primary objective, each subject was briefed as follows:

> The primary task is to track the target as accurately as possible, keeping the center of the target within the "diamond" symbol. It is not essential that your "wings" are level. The secondary objective is to keep the "wings" as level as possible.

While maintaining the tracking error below a criterion value, roll difficulty increases. Conversely, when the tracking error exceeds the criterion value, the roll difficulty decreases. The experiment is terminated if the display limits are exceeded.

Figure 3. Primary Tracking Task with Secondary Loading Task



Seven different systems typifying different handling qualities are used:

| SYSTEM | ASSOCIATED HANDLING QUALITIES |
|---|---|
| $K/s$ | Rate Control: This system approximates heavily damped automobile turn control at a fixed speed. |
| $K/s(s+1)$ | Rate Control with Lag. |
| $K/s(s+2)$ | Rate Control with Lag. |
| $K/s(s+4)$ | Rate Control with Lag. |
| $K/[s^2 + 2(0.7)7.8s + 7.8^2]$ (Poly1) | Oscillatory Response: System is lightly damped with natural frequency less than the bandwidth of the tracking target pitch. |
| $K/[s^2 + 2(0.7)16s + 16^2]$ (Poly2) | Oscillatory Response: System is lightly damped with natural frequency greater than the bandwidth of the tracking target pitch. |
| $K/s^2$ | Acceleration Control: Acceleration is commanded by control input. Rate change depends on duration of control input. |

These seven different systems are further modified via gain and screen vertical display scale selection to model secondary variations in handling qualities. The combinations of systems and gains used are listed in Table 2.

**Table 2.** System and Gain Configurations

| SYSTEM | CONTROLLED ELEMENT GAIN | | | ADDITIONAL GAINS |
|---|---|---|---|---|
| | LOW | NOMINAL | HIGH | |
| $K/s$ | .293 | .586 | 5.86 | 1.17, 8.38 |
| $K/s(s+1)$ | 17.6 | 35.2 | 58.6 | 83.8 |
| $K/s(s+2)$ | 17.6 | 35.2 | 58.6 | 8.38, 21.5 |
| $K/s(s+4)$ | 17.6 | 35.2 | 58.6 | 83.8, 117 |
| Poly1 | 21.5 | 35.2 | 83.8 | 17.6, 58.6 |
| Poly2 | - | 35.2 | 83.8 | 215 |
| $K/s^2$ | .293 | .586 | 1.17 | - |

### B. Procedure

The experiment was conducted at the U.S. Naval Test Pilot School, Naval Air Test Center, Patuxent River, Maryland during August and September 1990.

A total of fifteen pilots participated in the experiment. Experimental subjects are graduates of the Naval Test Pilot School. All except one are assigned to aviation duties at the Naval Air Test Center. A questionnaire provided details regarding each person's background. This data is summarized in Appendix A.

Each pilot was briefed regarding the objectives of the experiment and the general characteristics of each configuration. As noted above, during the initial briefing, and prior to most runs, each subject was instructed and reminded that the primary task was to minimize pitch tracking error (i.e., to keep the target within the "diamond" symbol). The display was then selected and symbols defined. Each pilot positioned the hand controller for convenient operation.

A training segment was provided at the start of each session, based on the K/s system with gain .586 and "normal" vertical display scale. The pilot performed this "training" task until he felt comfortable and confident; both one axis (pitch) and two axis (pitch and roll) were demonstrated.

Following the training segment, each pilot executed a different sequence of runs where a run is determined by the particular configuration of system, gain, vertical display scale of the screen, and the bandwidth controlling the pitch of the tracking target. In addition, runs were made either with or without the secondary loading task. The run sequence was devised to expose each pilot to a range of configurations with the associated variations in characteristics of handling qualities. The total number of runs flown by each pilot ranged from 10 to 17; most flew 10 to 12. The exact sequence of runs executed by each pilot is listed in Appendix B.

Following each run, using the scoring procedure followed by the Naval Test Pilot School, the pilot was asked to comment on and evaluate the handling qualities typified by the system. That scoring procedure is based on the Cooper-Harper evaluation scheme and is described in Appendix C. During the initial briefing period, this procedure was reviewed and discussed with each pilot.

### C. Experimental Measurements

Measurements of pilot response and pitch/roll state were sampled and recorded at approximately fixed time intervals of 0.05 seconds for approximately 200 seconds. The specific variables recorded are listed in Table 3. In addition, the system, gain, vertical display scale, and bandwidth selected for the run were recorded. Following the run, pilot comments and subjective scores were recorded in the header of the data record.

**Table 3.** Variables Sampled and Recorded During Each Run for Each Pilot

| Channel | Variable |
|---|---|
| 1 | $\theta_c$ Pitch, Tracking Target |
| 2 | $\theta_e$ Pitch Tracking Error, Primary Task |
| 3 | $\phi$ Roll Error, Secondary Task |
| 4 | $\delta_e$ Pitch Control Input |
| 5 | $\delta_a$ Roll Control Input |
| 6 | $\lambda_s$ Secondary Task Parameter |
| 7 | $\Lambda_s$ Secondary Task Parameter |

The simulation software is designed to sample all variables at 0.05 second intervals. Attempts were made to use a smaller sampling interval but this is the fastest rate at which the IRIS 3000 hardware would respond consistently. Even at this time interval, the sampling is interrupted at times for up to 1.00 second leaving "gaps" in the sampled data. Instead of increasing the sampling interval to eliminate this irregularity in the sampling, a sampling rate of 0.05 second was used and the data subsequently interpolated to compensate for the gaps. The data was interpolated using cubic spline interpolation on each variable separately which slightly distorted the dynamic relationship between variables; since the width of the gaps were not necessarily multiples of 0.05, the interpolation essentially resampled the data instead of merely filling in the "gaps". The effect of this distortion was investigated and was found not to significantly affect subsequent analyses.

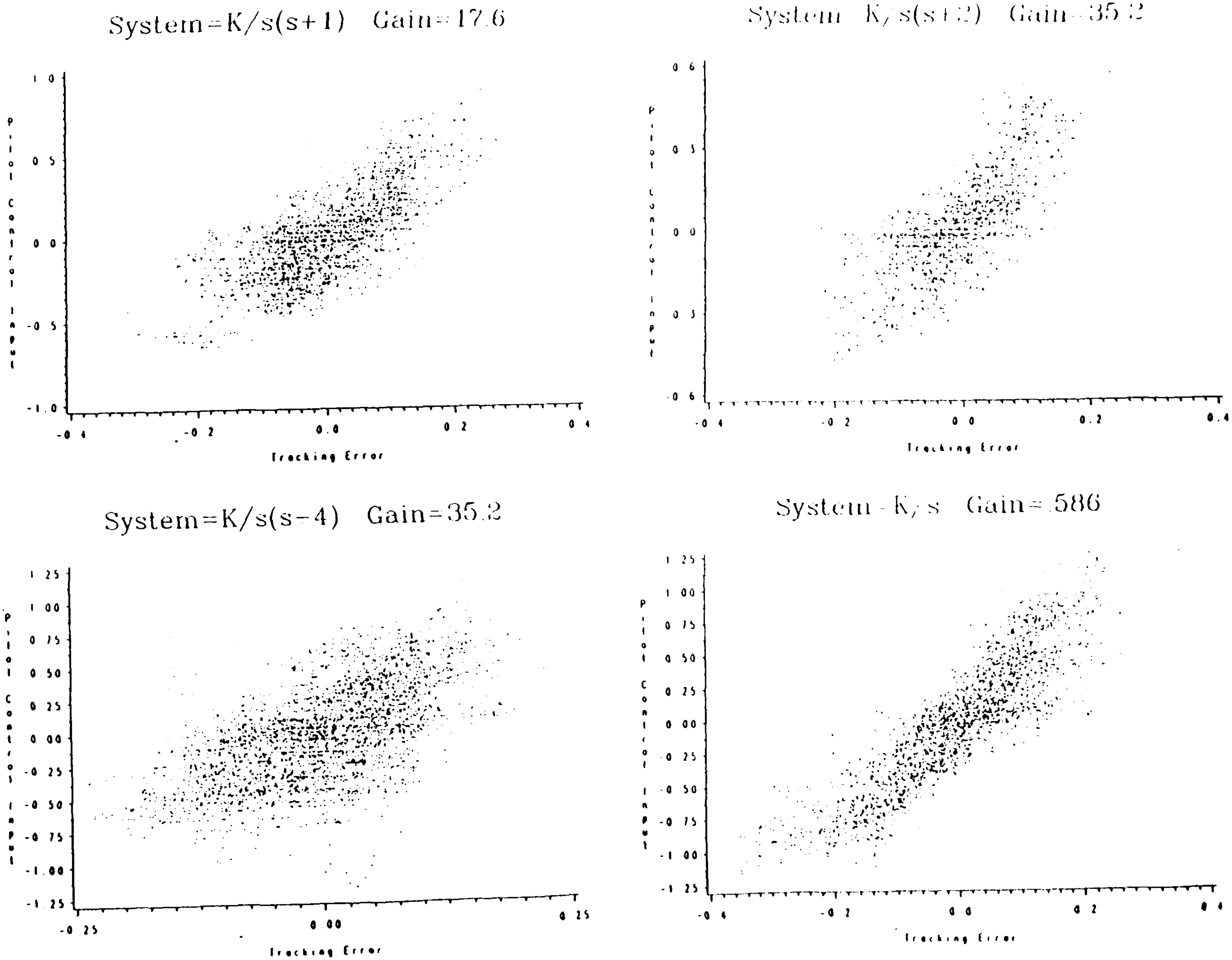
## D. Lead and Lag

Simulations of this type in which a pilot interacts with a computer-driven simulation often result in a time differential between the pilot's stimulation (i.e., the current state of the pitch and roll) and the pilot's reaction (i.e., the inputted control to adjust the pitch and roll). The term lag refers to a pilot's reaction being subsequent to the stimulation; the term lead refers to a pilot's reaction preceding the stimulation as might occur if the pilot starts anticipating the direction of movement of the target being tracked.

Existence of either lead or lag was investigated by computing and plotting the cross-correlation function between $\theta_e$ and $\delta_e$ and between $\phi$ and $\delta_a$ for each run made by each pilot. The cross-correlation function between $\theta_e$ and $\delta_e$ revealed a fairly constant lag of 0.40 seconds over both pilot and system configurations. At this lag, the statistical correlation coefficient between $\theta_e$ and $\delta_e$ is typically 0.85 for the "simpler" systems like K/s and K/s(s+a) degrading to 0.55 for the most difficult system $K/s^2$. No substantial evidence of a lead was found in any system configuration or pilot although some isolated system configurations and pilot combinations did show large correlations at negative lag times. The cross-correlation function between $\phi$ and $\delta_a$ showed no evidence of either a lead or lag.

In view of these findings, the inputted control measurements were shifted 0.40 seconds (8 units of 0.05 seconds each) backward relative to the pitch/roll state sequence when investigating relationships between pilot inputted control and the resulting state of pitch and roll. Scatter plots of $\delta_e$ against $\theta_e$, with $\theta_e$ lagged by 0.40 seconds are presented in Figure 4 for selected experimental runs. The general nonlinearity of the scatter plots of $\delta_e$ against lagged $\theta_e$ suggests the need to model the control law as nonlinear.

Figure 4. Scatter Plots of $\delta_e$ against $\theta_e$, with $\theta_e$ Lagged by 0.40 Seconds

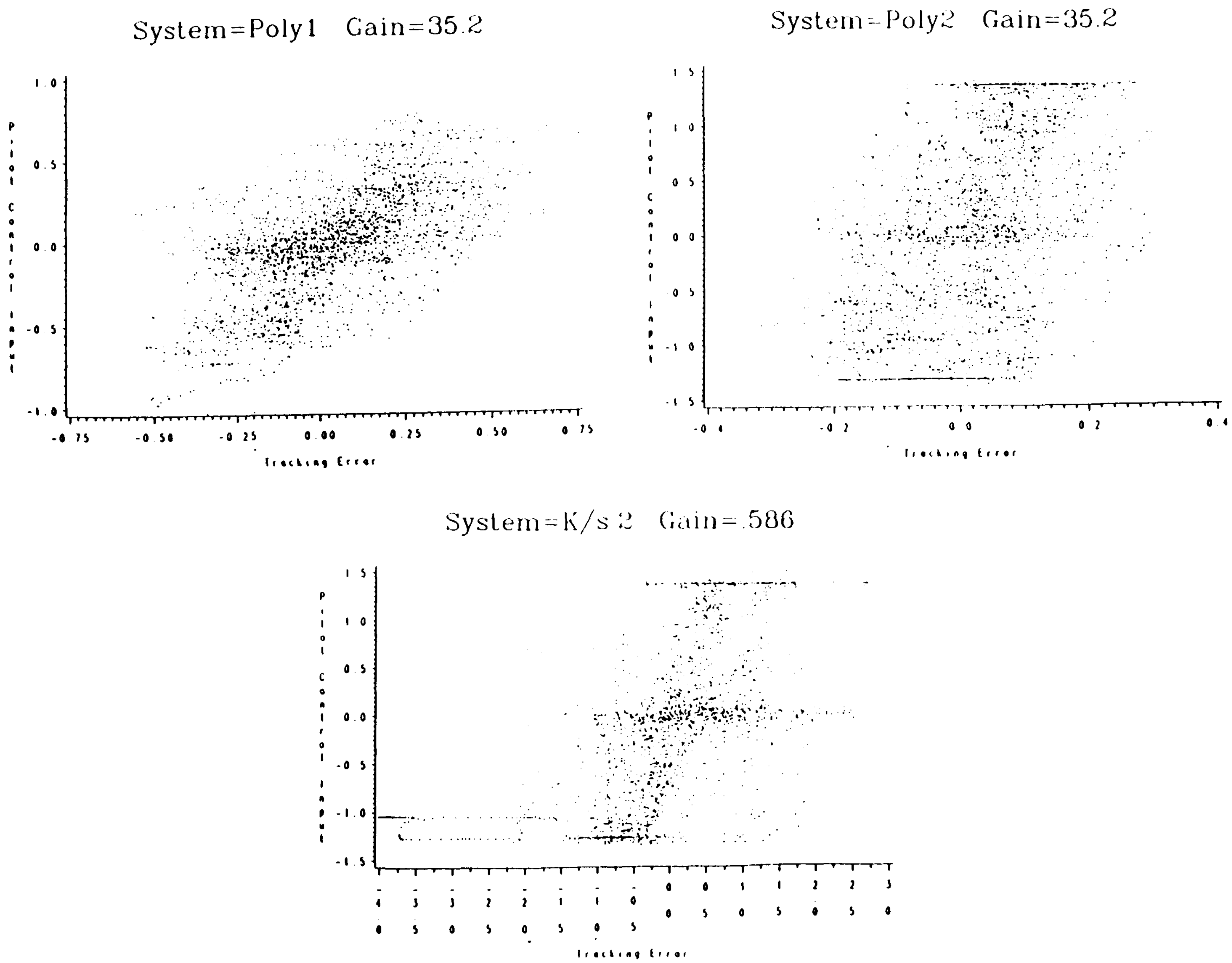

System=K/s(s+1) Gain=17.6

System=K/s(s+2) Gain=35.2

System=K/s(s-4) Gain=35.2

System=K/s Gain=586

Figure 4. (Continued) Scatter Plots of $\delta_e$ against $\theta_e$, with $\theta_e$ Lagged by 0.40 Seconds



# SECTION III
# PILOT RESPONSE MEASUREMENTS

## A. Overview

Each run measures pilot response to a system during a tracking task. Maintaining a homogeneous level of difficulty for the task over the run yields a large data set of pilot response measurements under similar handling qualities. In statistical terms, each run provides a data set with a large number of replicate measurements. If the pilot is determining and applying control inputs consistently, then statistical analysis of pilot response measurements can be used to construct an estimated pilot control law. For this project, the estimated control law is a mathematical formula for calculating a control input for the pitch dimension in terms of the history of simulator states. The following paragraphs describe how the pitch control input law is defined by a logistic function of tracking error.

Then, based on the estimated control law, two classes of performance measures are defined:

> The process of statistical estimation yields fit and error measures associated with the estimation; such measures include a regression mean square (RMS), an error mean square (EMS), and an R-square[1] (RSQ).
>
> The mathematical formula for pitch control is used to provide pitch control input in a closed loop tracking task simulation for the same system flown by the pilot. Mean absolute tracking error (MEAN) is calculated over this closed loop simulation run, and this quantity provides a measure of tracking accuracy for the estimated control law.

The authors' previous work (Baldwin and Gantz, 1983, 1984) indicated that RMS, EMS, RSQ and MEAN may provide useful measures for the evaluation of pilot performance. This section of the report describes how these measures are defined and calculated for each experimental run in the project. Further, the specific approach to statistical modeling of pilot control yields additional measures of potential value for the quantification of pilot performance for use in evaluating handling qualities and understanding Cooper-Harper ratings.

[1] The equation for the R-square value is given by

RSQ = 1 - (error sum of squares)/(total sum of squares).

## B. Control Law Modeling

For each experimental run, the pilot's control inputs are assumed to represent an "optimal" solution to the tracking task in the sense of optimal control theory. For each run, a control law describing the relationship between measurements observable by the pilot and the pilot's pitch control input is estimated statistically. The estimate of the pilot control law provides a rule for optimally completing the tracking task. Statistical analysis of the correlation between pilot pitch control inputs and observaole measurements shows that $\theta_e$, the tracking error, is the only measurement correlated with pilot control inputs. Regression analysis is used to estimate the relationship between tracking error and pilot control inputs.

In the Linear-Quadratic-Gaussian optimal control problem, an optimal control law is linear. The system designed for these experiments is not linear; however, as a first cut at modeling, linear regression was used to relate the pilot's observed control inputs to tracking error. Linear regression modeling did not properly fit the experimental data. Examples of linear regression fits and the associated residual plots for a linear regression fit are presented in Figure 5. The residuals in the figure have a pattern suggesting that a better fit could be obtained by estimating a type of nonlinear saturated control law. An example of a saturated control law is shown in Figure 6.

**Figure 6.** A Saturated Control Law

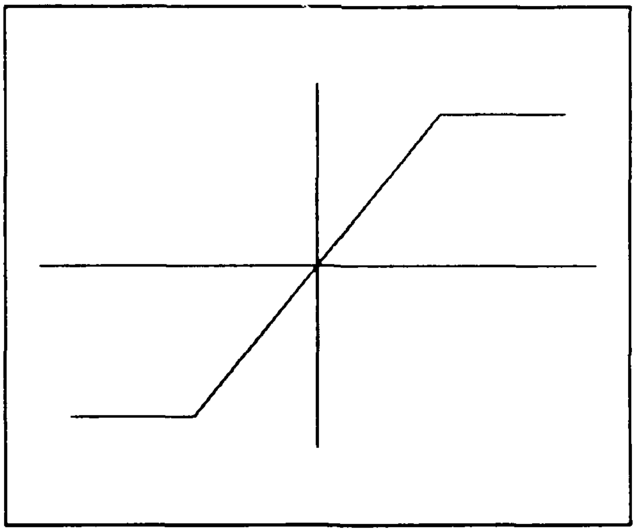
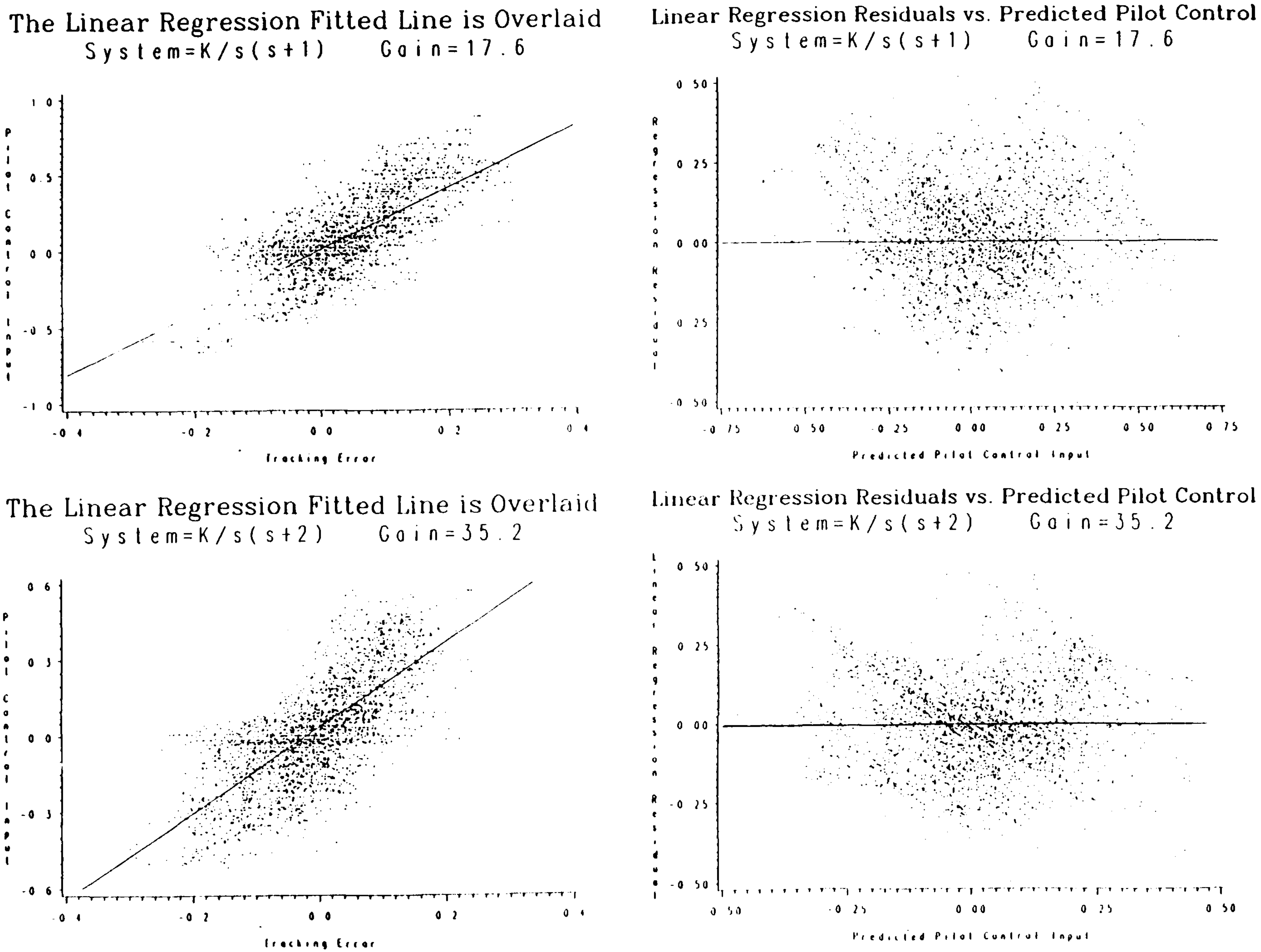


**Figure 5. Linear Regression Fits and Residual Plots ($\delta_e$ against Lagged $\theta_e$)**



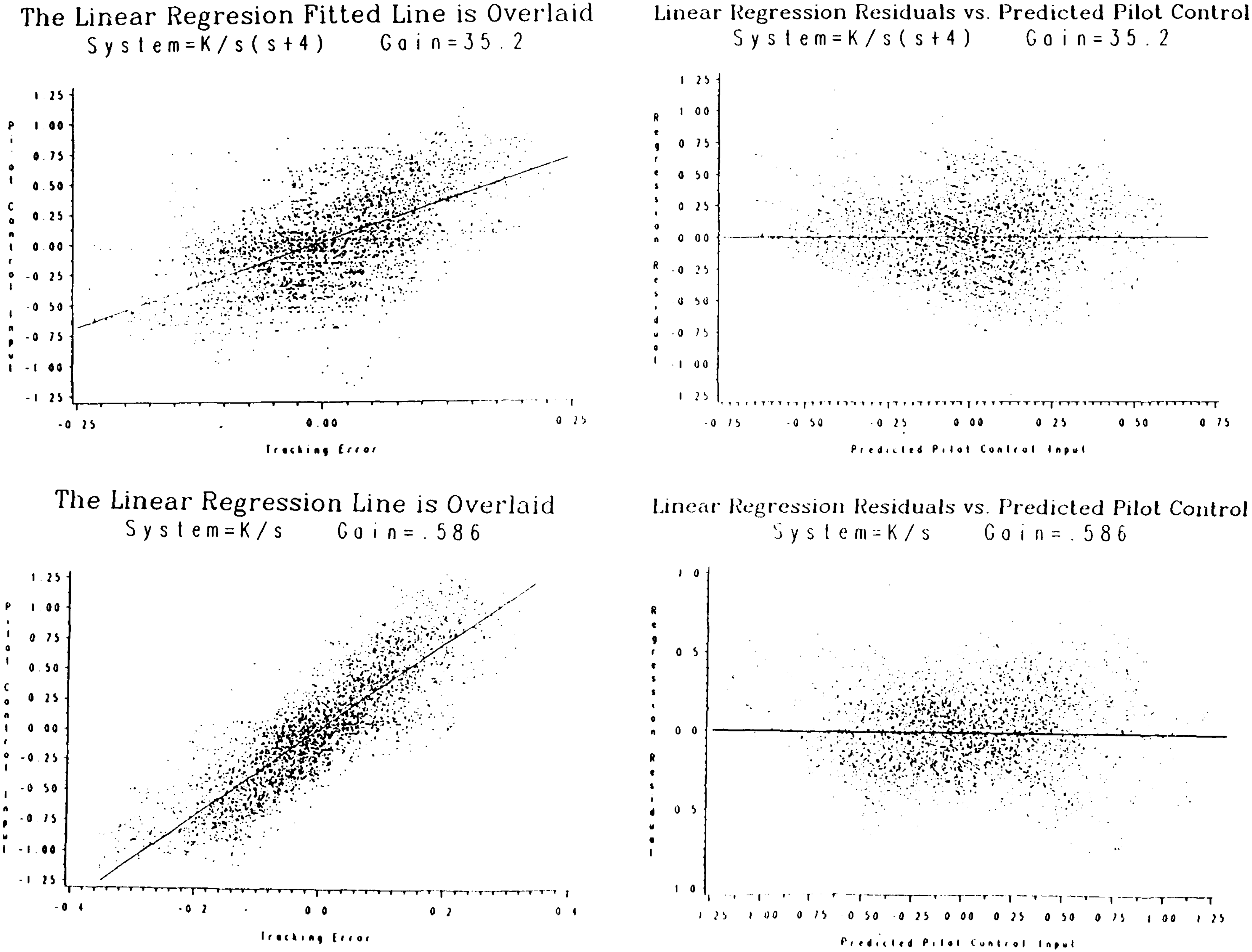
The Linear Regression Fitted Line is Overlaid
System=K/s(s+1) Gain=17.6

Linear Regression Residuals vs. Predicted Pilot Control
System=K/s(s+1) Gain=17.6

The Linear Regression Fitted Line is Overlaid
System=K/s(s+2) Gain=35.2

Linear Regression Residuals vs. Predicted Pilot Control
System=K/s(s+2) Gain=35.2

**Figure 5. (Continued) Linear Regression Fits and Residual Plots ($\delta_e$ against Lagged $\theta_e$)**



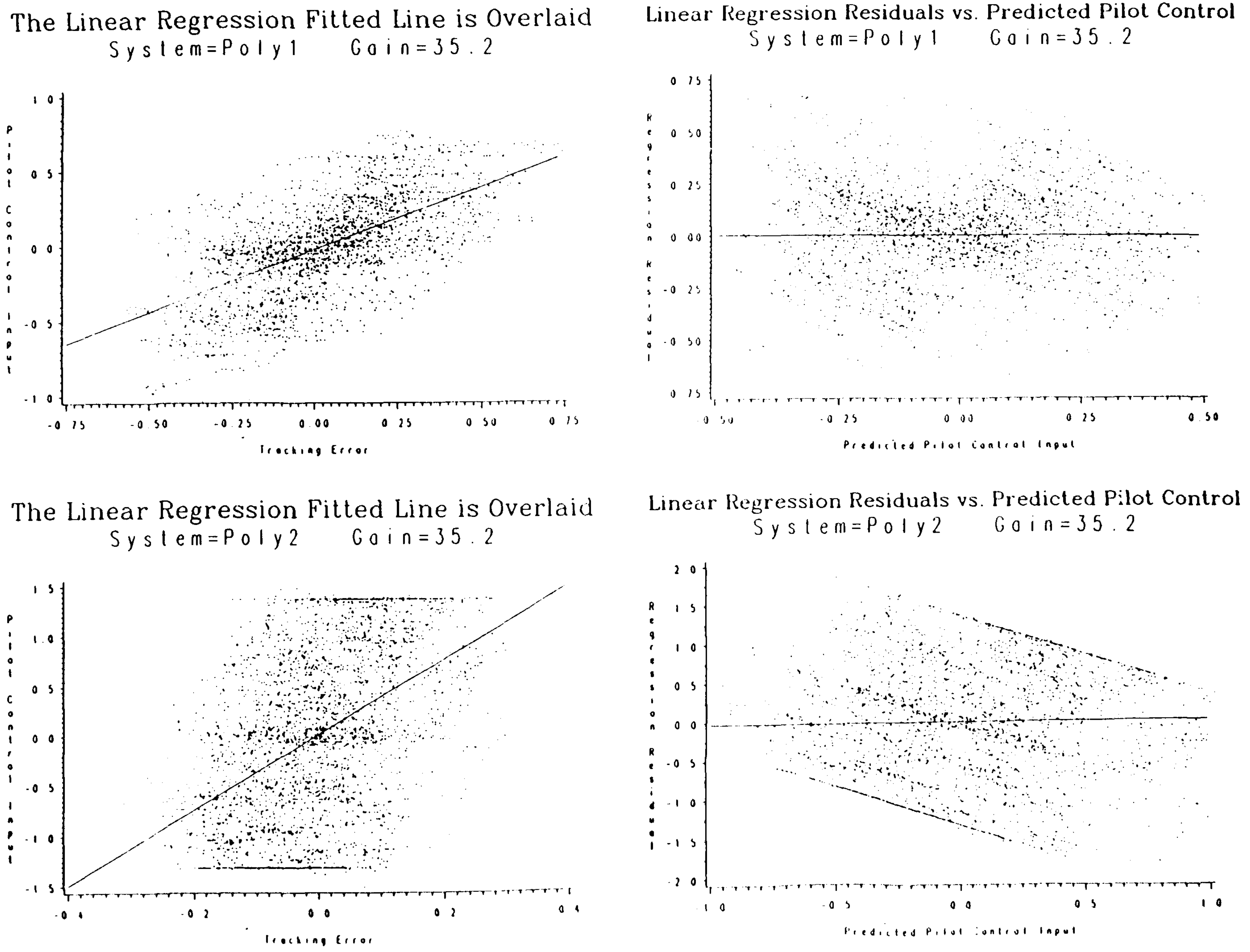
The Linear Regresion Fitted Line is Overlaid
System=K/s(s+4) Gain=35.2

Linear Regression Residuals vs. Predicted Pilot Control
System=K/s(s+4) Gain=35.2

The Linear Regression Line is Overlaid
System=K/s Gain=.586

Linear Regression Residuals vs. Predicted Pilot Control
System=K/s Gain=.586

## Figure 5. (Continued) Linear Regression Fits and Residual Plots ($\delta_e$ against Lagged $\theta_e$)

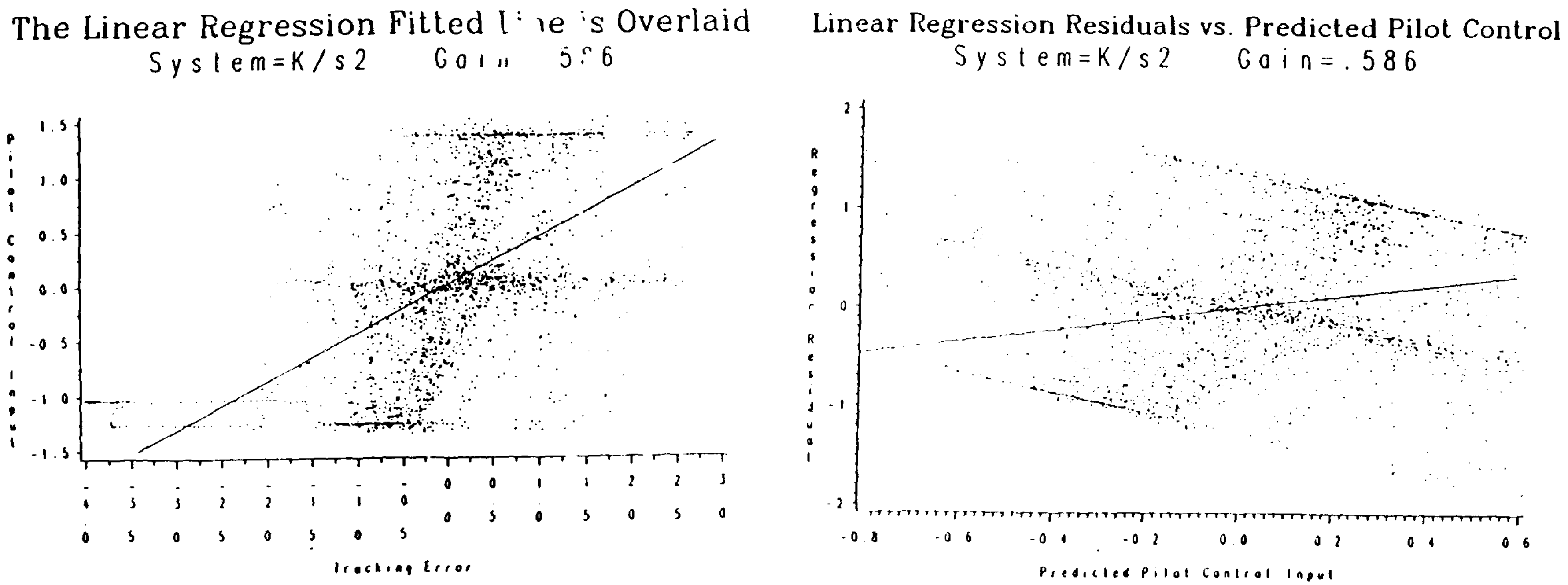

The Linear Regression Fitted Line is Overlaid
System=Poly1 Gain=35.2

Linear Regression Residuals vs. Predicted Pilot Control
System=Poly1 Gain=35.2

The Linear Regression Fitted Line is Overlaid
System=Poly2 Gain=35.2

Linear Regression Residuals vs. Predicted Pilot Control
System=Poly2 Gain=35.2

**Figure 5. (Co‑tinued) Linear Regression Fits and Residual Plots ($\delta_e$ against Lagged $\theta_e$)**



It was judged that a logistic curve representation of a saturated control law is flexible enough to fit the data. The logistic curve is a four parameter curve. The formula for the logistic curve in terms of the four parameters, $B_1$-$B_4$, is:

$$\delta_e = B_1 + \frac{B_2}{1 + \exp(-B_3 - B_4\theta_e)}$$

Nonlinear regression analysis was used to estimate the parameters $B_1$-$B_4$ that best fit the logistic curve to test data. The nonlinear regression estimation was accomplished through the PROC NLIN procedure of the SAS/STAT system (1985). This procedure was run on each of the experimental data sets. Hence, estimated parameters for a logistic shaped control law were calculated for each experimental run.

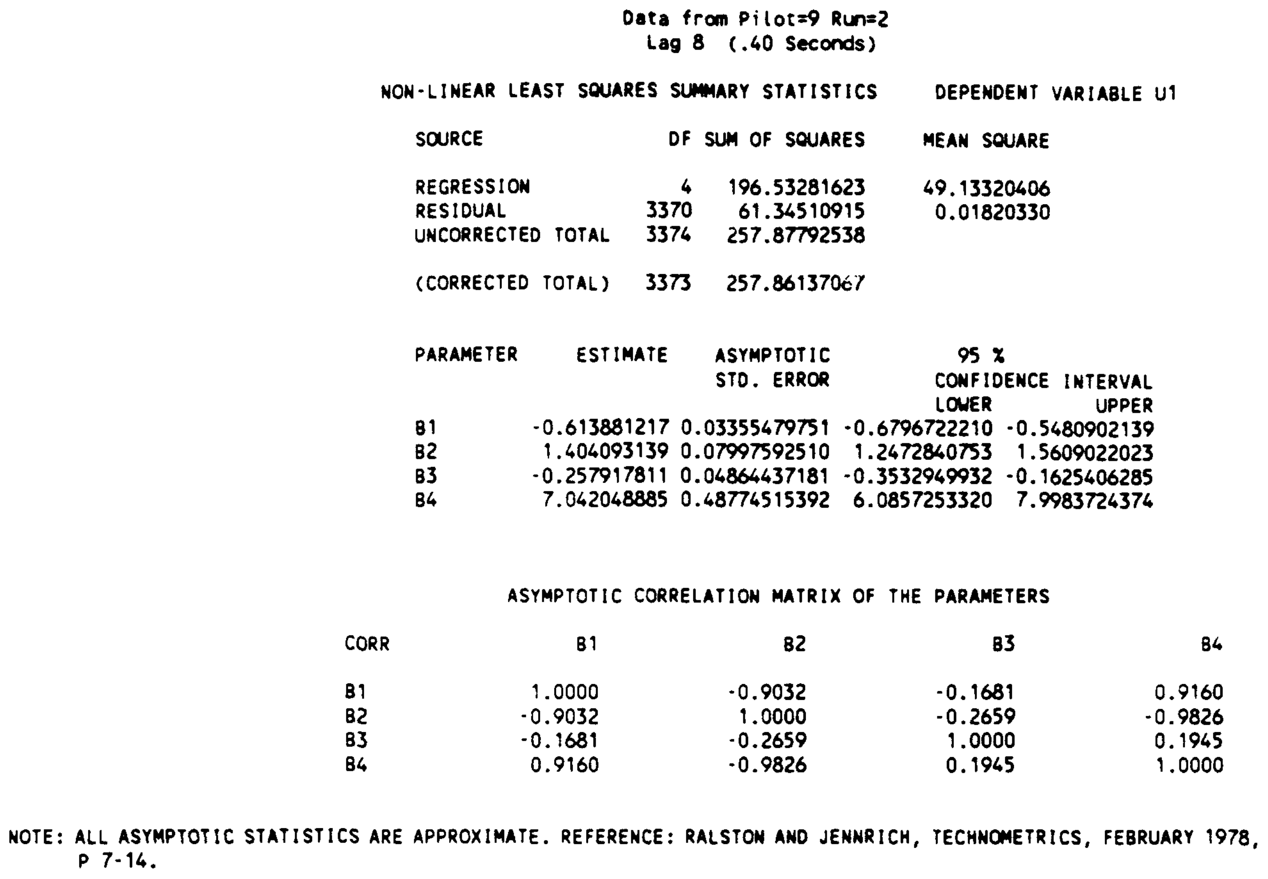
It was reported above that the highest statistical correlation between tracking error and pilot control inputs was for tracking error lagged by 0.40 seconds. The nonlinear regression fitting was performed between the pilot pitch control input ($\delta_e$) as the dependent variable and the lagged tracking error (lagged $\theta_e$) as the independent variable. Additional measurement variables (for example, the derivative of tracking error) were added as independent variables with no improvement of fit. Hence, lagged tracking error remained the single independent variable in the analysis.

A sample PROC NLIN output is presented in Figure 7. The PROC NLIN procedure iteratively fits a logistic curve to the data. Initial parameters for the curve are provided to the procedure. A common set of parameter initializations worked well for many of the experimental data runs. However, about twenty percent of the runs required customized initial parameters to get a satisfactory curve fit to the data. Once the NLIN procedure has judged that the parameters are providing an optimal fit to the data, regression statistics are calculated for the fit. These regression statistics include the regression mean square (RMS), the error mean square (EMS), and the R-square (RSQ) value as well as the parameter estimates of $B_1$-$B_4$ with associated errors and significance levels. Examples of curve fits and residual plots for the logistic curve fitting are presented in Figure 8. The nonlinear fit has removed the nonlinear pattern in the residual plots for a linear fit shown previously in Figure 5.

## Figure 7. PROC NLIN Output

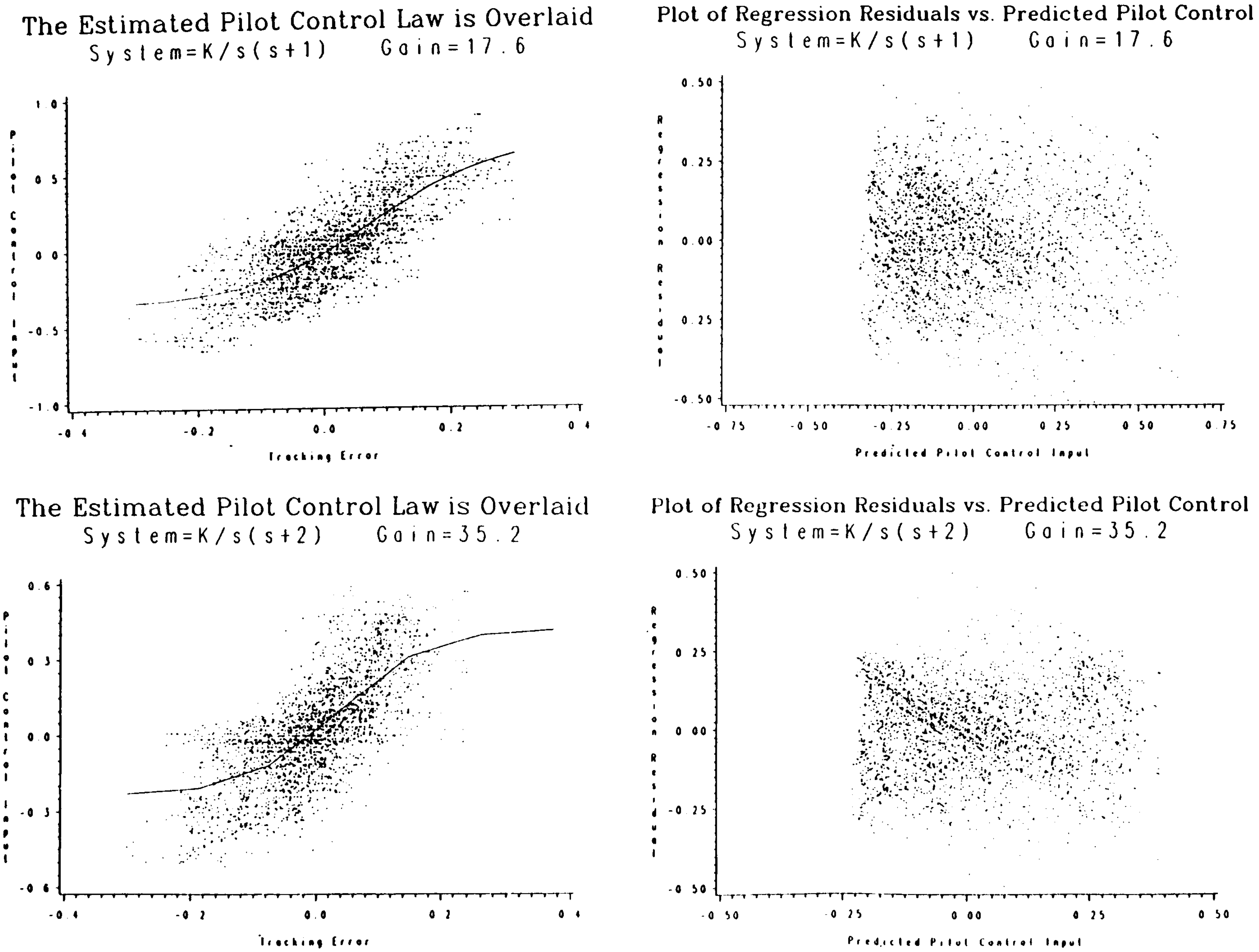
```
                              Data from Pilot=9 Run=2
                               Lag 8  (.40 Seconds)

                       NON-LINEAR LEAST SQUARES ITERATIVE PHASE

                    DEPENDENT VARIABLE: U1       METHOD: DUD

ITERATION         B1            B2              B3            B4            RESIDUAL SS

   -5          -3.86068    8.829203000      -0.25768    1.119963000     68.145056352365
   -4         -4.246748    8.829203000      -0.25768    1.119963000    622.559524769012
   -3          -3.86068    9.712123300      -0.25768    1.119963000    523.601524124247
   -2          -3.86068    8.829203000     -0.283448    1.119963000     85.772137354046
   -1          -3.86068    8.829203000      -0.25768    1.231959300     76.450579455773
    0          -3.86068    8.829203000      -0.25768    1.119963000     68.145056352365
    1         -3.573701    8.089688125    -0.2413057    1.211122397     68.060153772589
    2         -3.499812    7.903180290    -0.2375287    1.236200553     68.034855791150
    3         -2.637192    5.808340559    -0.1972555    1.604880490     66.477651380366
    4         -2.060521    4.461694107    -0.1737488    1.935205584     66.130343765427
    5         -2.185615    4.768312950    -0.1796189    1.892703196     65.000106994184
    6         -2.189552    4.777993529    -0.1798368    1.892508917     64.997714943013
    7         -2.212332    4.837120629    -0.1811999    1.890984211     64.854285262640
    8         -1.890682    4.105909872    -0.1715656    2.164377374     64.427292444172
    9         -1.650716    3.568311224    -0.1659069    2.413992186     64.395460216124
   10         -1.754383    3.817137836    -0.1718359    2.413980018     64.240918232505
   11         -1.704392    3.697794433    -0.1690426    2.416868671     64.041072455446
   12         -1.711858    3.716059926    -0.1695447    2.418573246     64.037186416494
   13         -1.549999    3.359717456    -0.1674213    2.631936140     63.735202315165
   14         -1.434775    3.109279984    -0.1670841    2.825893934     63.540263654103
   15         -1.442596    3.127450301    -0.1676435    2.832921703     63.526494395540
   16         -1.443881    3.130416058    -0.1677149    2.833247555     63.525964329067
   17         -1.445298    3.133696481    -0.1677962    2.833631860     63.525697947657
   18         -1.356649    2.943058410    -0.1690755    3.007431052     63.373547647801
   19         -1.282128    2.782973231    -0.1703638    3.171524626     63.248496566157
   20        -0.6128807    1.393119670    -0.2245733    7.709795065     62.471279781326
   21        -0.5813523    1.325169527    -0.2247343    7.741325747     61.697460419705
   22        -0.5758633    1.313125879    -0.2246048    7.737304925     61.650631882659
   23        -0.5708183    1.302049197    -0.2244792    7.733130747     61.639785971308
   24        -0.6295553    1.428249130    -0.2195546    6.769830941     61.530250211234
   25        -0.6302112    1.429421478    -0.2196963    6.835025113     61.500343355426
   26        -0.6293308    1.427416609    -0.2196276    6.857264828     61.500041767061
   27        -0.6295204    1.429513389    -0.2249049    6.908113084     61.467060283092
   28        -0.6037107    1.384128041    -0.2603244    7.016222138     61.405525825519
   29        -0.6128344    1.403148003    -0.2607275    7.051195904     61.345332078636
   30        -0.6133932    1.404200962    -0.2596544    7.043094324     61.345134383675
   31        -0.6133788    1.404091989    -0.2595751    7.043841536     61.345132740936
   32         -0.613574    1.404415245    -0.2593332    7.041202298     61.345125717978
   33        -0.6138838    1.404099735    -0.2579187    7.042014718     61.345109154069
   34        -0.6138812    1.404093139    -0.2579178    7.042048885     61.345109153317

NOTE: CONVERGENCE CRITERION MET.
```

## Figure 7. PROC NLIN Output (Continued)



```
                                  Data from Pilot=9 Run=2
                                   Lag 8  (.40 Seconds)

               NON-LINEAR LEAST SQUARES SUMMARY STATISTICS     DEPENDENT VARIABLE U1

                  SOURCE                 DF  SUM OF SQUARES      MEAN SQUARE

                  REGRESSION              4    196.53281623      49.13320406
                  RESIDUAL             3370     61.34510915       0.01820330
                  UNCORRECTED TOTAL    3374    257.87792538

                  (CORRECTED TOTAL)    3373    257.86137067

                  PARAMETER      ESTIMATE      ASYMPTOTIC             95 %
                                               STD. ERROR        CONFIDENCE INTERVAL
                                                                 LOWER          UPPER
                  B1         -0.613881217 0.03355479751 -0.6796722210 -0.5480902139
                  B2          1.404093139 0.07997592510  1.2472840753  1.5609022023
                  B3         -0.257917811 0.04864437181 -0.3532949932 -0.1625406285
                  B4          7.042048885 0.48774515392  6.0857253320  7.9983724374


                             ASYMPTOTIC CORRELATION MATRIX OF THE PARAMETERS

             CORR               B1               B2               B3               B4

             B1             1.0000          -0.9032          -0.1681           0.9160
             B2            -0.9032           1.0000          -0.2659          -0.9826
             B3            -0.1681          -0.2659           1.0000           0.1945
             B4             0.9160          -0.9826           0.1945           1.0000

NOTE: ALL ASYMPTOTIC STATISTICS ARE APPROXIMATE. REFERENCE: RALSTON AND JENNRICH, TECHNOMETRICS, FEBRUARY 1978,
      P 7-14.
```

**Figure 8. Nonlinear (Logistic Curve) Regression Fits and Residual Plots ($\delta_e$ against Lagged $\theta_e$)**

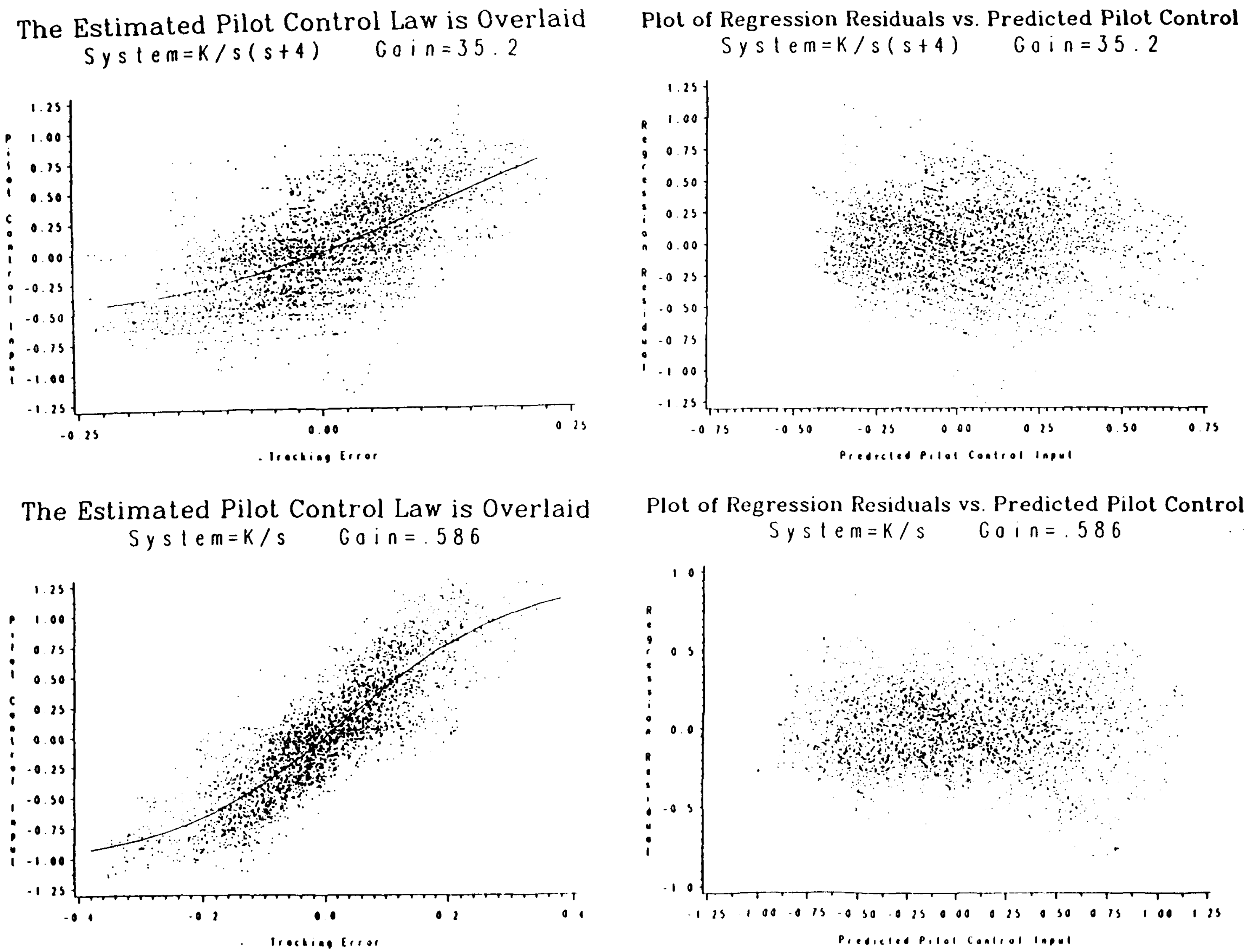

The Estimated Pilot Control Law is Overlaid
System=K/s(s+1)   Gain=17.6

Plot of Regression Residuals vs. Predicted Pilot Control
System=K/s(s+1)   Gain=17.6

The Estimated Pilot Control Law is Overlaid
System=K/s(s+2)   Gain=35.2

Plot of Regression Residuals vs. Predicted Pilot Control
System=K/s(s+2)   Gain=35.2

Figure 8. (Continued) Nonlinear (Logistic Curve) Regression Fits and Residual Plots ($\delta_e$ against Lagged $\theta_e$)



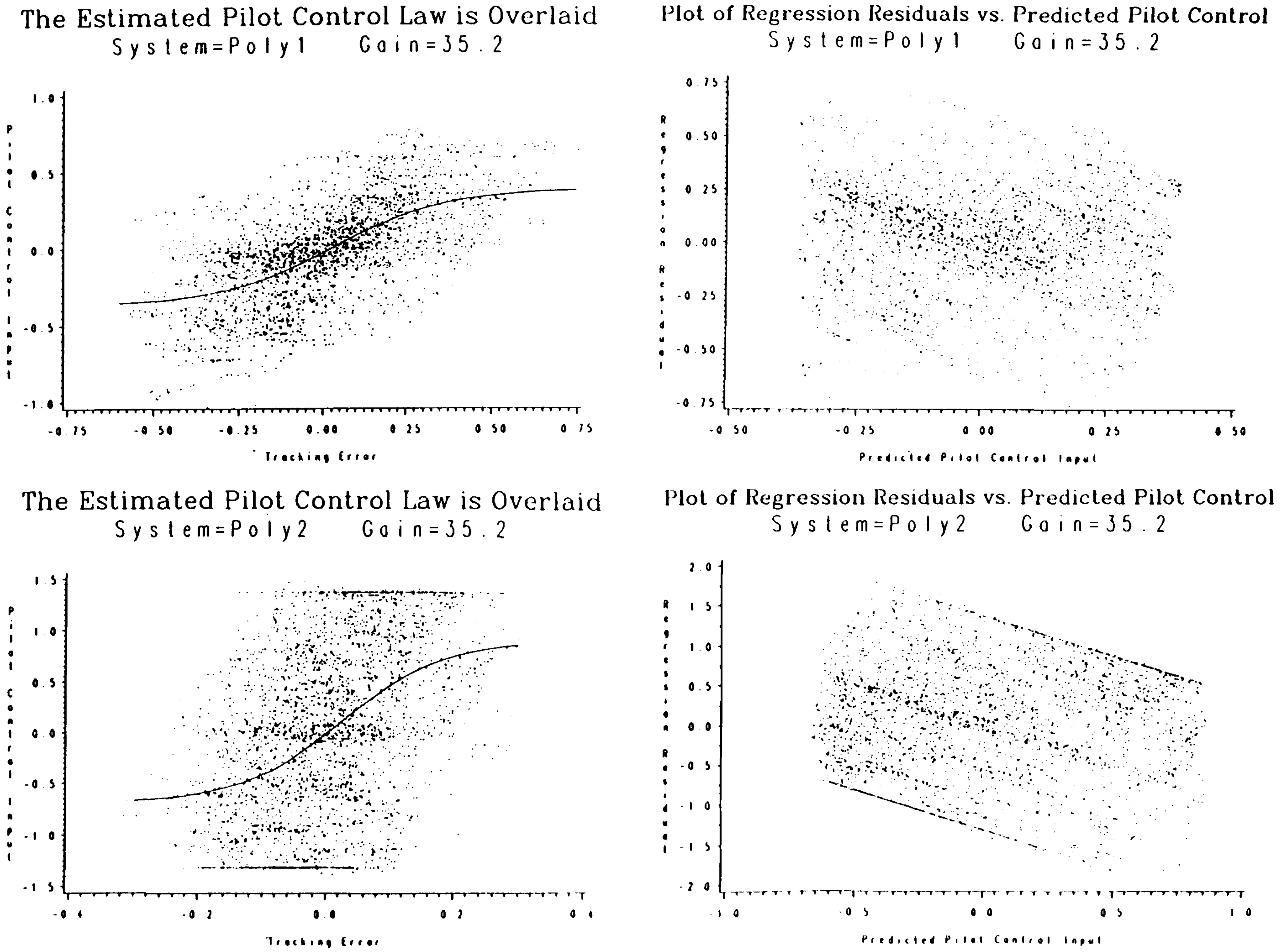
The Estimated Pilot Control Law is Overlaid
System=K/s(s+4) Gain=35.2

Plot of Regression Residuals vs. Predicted Pilot Control
System=K/s(s+4) Gain=35.2

The Estimated Pilot Control Law is Overlaid
System=K/s Gain=.586

Plot of Regression Residuals vs. Predicted Pilot Control
System=K/s Gain=.586

**Figure 8. (Continued) Nonlinear (Logistic Curve) Regression Fits and Residual Plots ($\delta_e$ against Lagged $\theta_e$)**



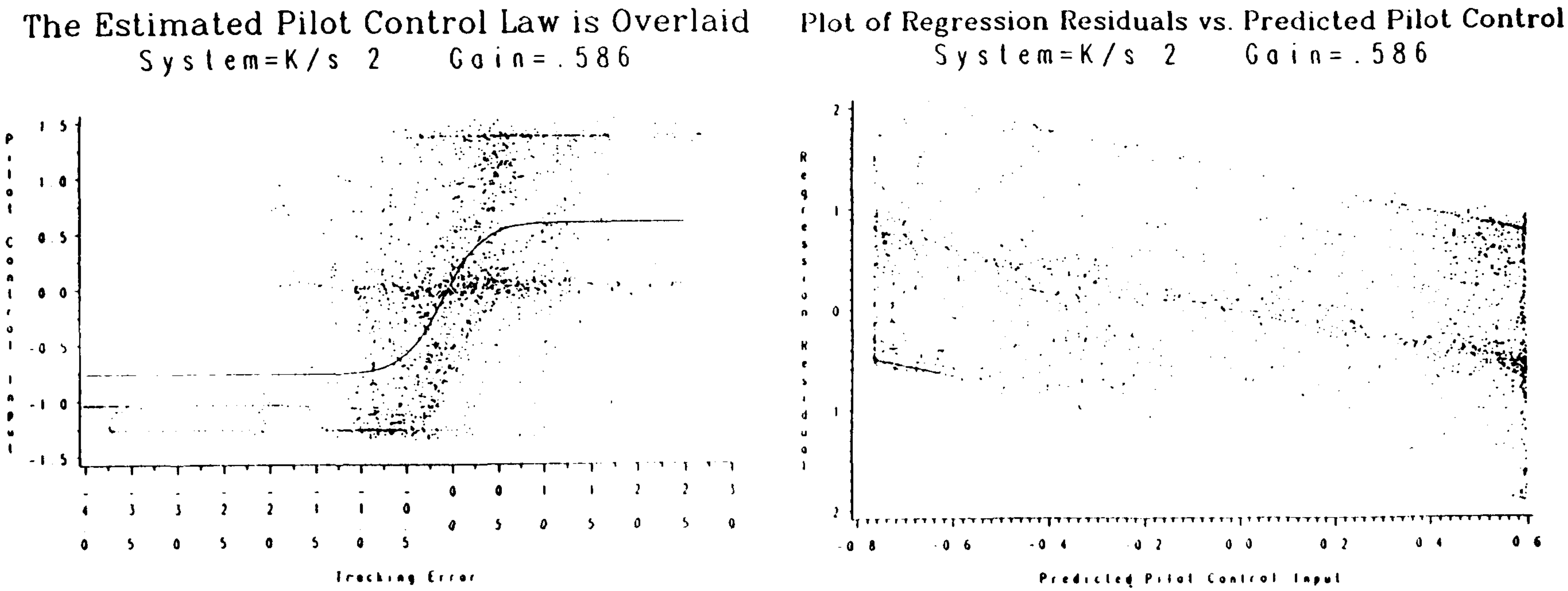
The Estimated Pilot Control Law is Overlaid
System=Poly1 Gain=35.2

Plot of Regression Residuals vs. Predicted Pilot Control
System=Poly1 Gain=35.2

The Estimated Pilot Control Law is Overlaid
System=Poly2 Gain=35.2

Plot of Regression Residuals vs. Predicted Pilot Control
System=Poly2 Gain=35.2

**Figure 8. (Continued) Nonlinear (Logistic Curve) Regression Fits and Residual Plots ($\delta_e$ against Lagged $\theta_e$)**



The Estimated Pilot Control Law is Overlaid
System=K/s 2 Gain=.586

Plot of Regression Residuals vs. Predicted Pilot Control
System=K/s 2 Gain=.586

## C. Primary Measures of Pilot Performance

Two primary categories of error measures were derived from estimating a control law for each experimental run. The definitions of these error measures are consistent with measures defined in the authors' earlier work (Baldwin and Gantz, 1983, 1984). The error measures are intended to provide objective information for evaluating system handling qualities and understanding Cooper-Harper ratings by condensing the measurements of pilot response to the system. The first category of error measures describes how well the pilot maintains a desired control law, and the second category of error measures describes how well the desired control law performs.

The diagrams in Figures 9 and 10 describe interpretatively how these two categories of error measures are calculated. The dotted box in Figure 9 represents the system defined earlier in Figure 3. The input signal, $\theta_c$, is differenced with the output signal, $\theta$, to get the tracking error, $\theta_e$. $\theta_e$ is fed back into the system definition to play a role in modulation of the stability parameter for the roll component. Further, $\theta_e$ is observed by the pilot for determination of the pitch control input, $\delta_e$, and also lagged and inputted to the estimated control law for calculation of the ideal control input, $\delta_i$. The squared difference of $\delta_e$ and $\delta_i$ is added to the sum of squared errors to calculate the mean square error for the nonlinear regression estimate. Actual calculation of the nonlinear regression mean square error is done through the SAS PROC NLIN procedure as described above. The error mean square (EMS) of the control law estimation is in the first category of performance measures referred to earlier in this paragraph; it tells how well the ideal control law is being implemented. Returning to the diagram, the output, $\theta$, of the pilot controlled system is differenced with the input signal, $\theta_c$, to accumulate the mean absolute tracking error of the pilot controlled system. The mean absolute tracking error (MEAN_RUN) is in the second category of performance error referred to earlier in this paragraph; it tells how well the pilot control inputs perform.

An additional second category measure which concerns how well the ideal control law performs is described in the diagram in Figure 10. Basically, this diagram has the ideal control law playing the role of the pilot. Graphically presented examples of how well the estimated control law tracks the input signal are found in Figure 11. The absolute (vertical) tracking error in these examples is averaged over the 180 second closed loop simulated run to get a simulated mean absolute tracking error (MEAN) for the estimated control law. The simulated mean absolute tracking error is a measure of performance for the ideal control law. This simulated mean absolute tracking error calculated by flying the estimated control law provides an error measure for the control law which is decoupled from the error measures resulting from the statistical estimation of the control law. The authors' earlier work (Baldwin and Gantz, 1983, 1984) indicated that the statistical error measures (EMS, RMS, and RSQ) together with the simulated mean absolute tracking error (MEAN) could provide a basis for evaluation of handling qualities.

**Figure 9.** Calculation of Performance Measures Describing Errors in Maintaining a Desired Control Law

The Dotted Box Represents the System Defined in Figure 3.

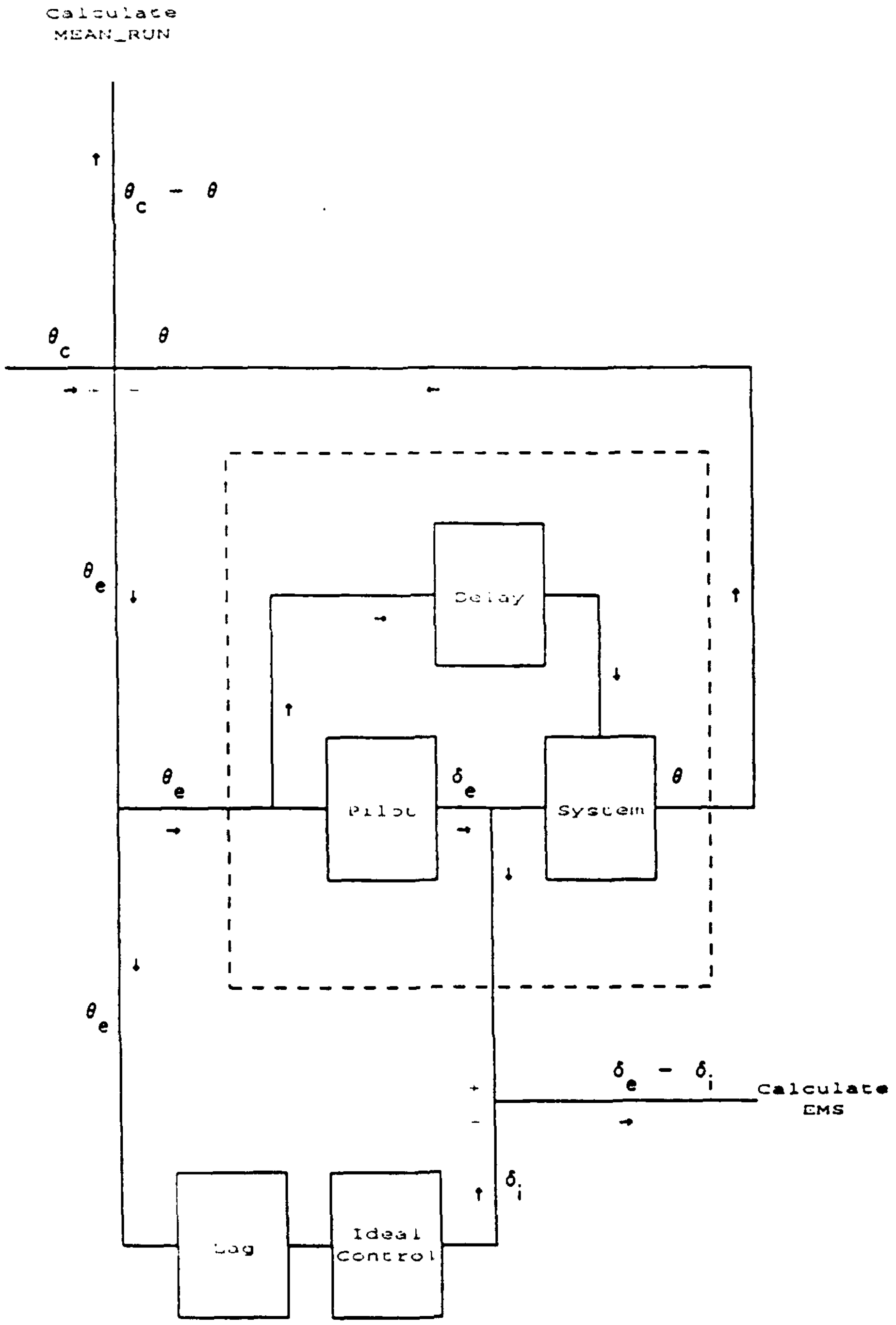


**Figure 10.** Calculation of the Per'formance Measure Describing How Accurately the Desired Control Law Performs the Primary Tracking Task

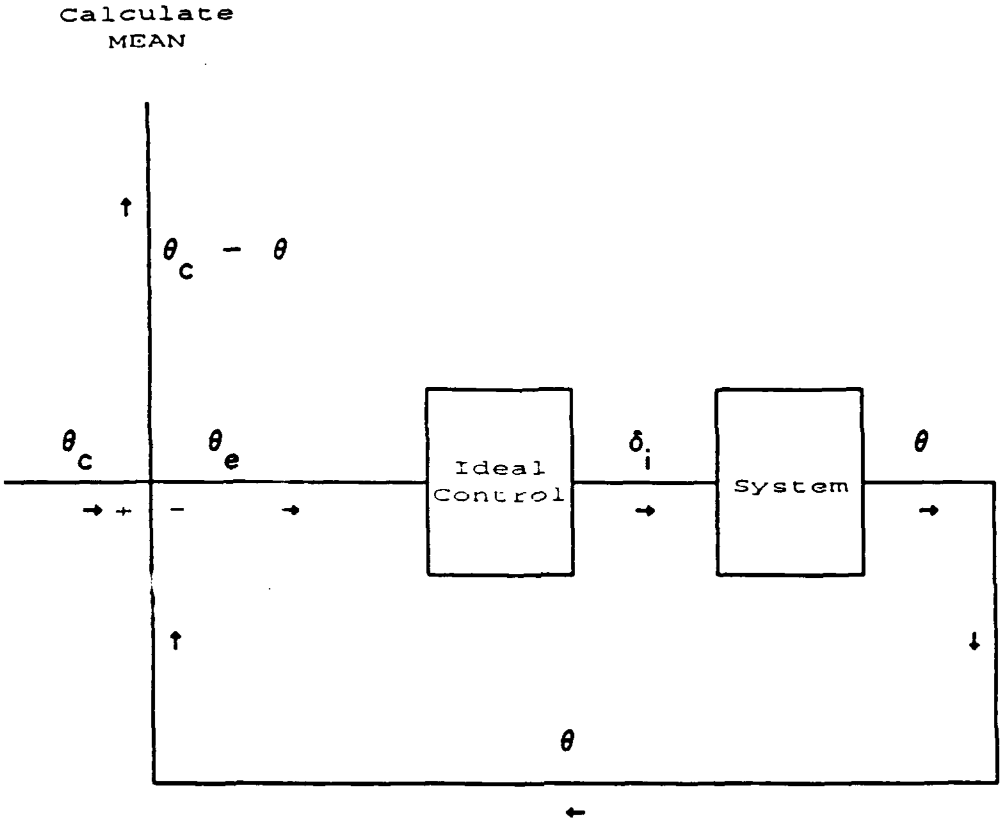
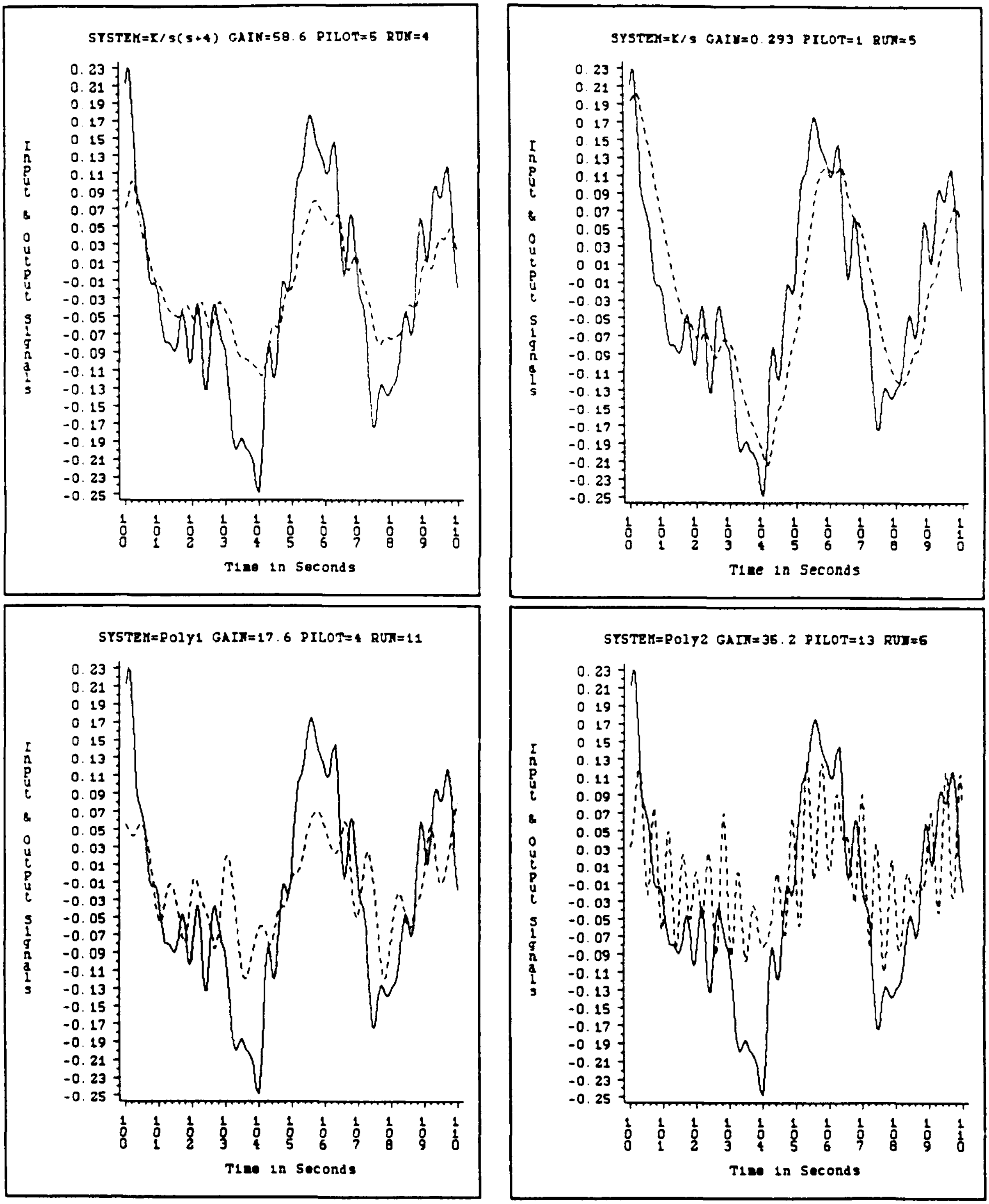


**Figure 11.** Tracking of the Input Signal by the Estimated Control Law
The Input Signal is Represented by a Solid Line
The Output Signal is Represented by a Dashed Line



## D. Secondary Measures of Pilot Performance

In addition to the performance measures defined in the preceding paragraphs, other objective measures were also calculated for each experimental run.

The secondary loop of the tracking task requires control of the roll axis. The roll control loop (see Figure 3) contains an instability parameter $\lambda_s$ which is a function of the pitch tracking error, $\theta_e$. $\lambda_s$ is modulated in order to keep an approximately steady level of difficulty for the tracking task during the run. Some examples of the variability of $\lambda_s$ during a run are presented in Figure 12.

It was assumed that periods during which pitch tracking error, $\theta_e$, was small required somewhat lessened pilot attention. During these periods, the value of $\lambda_s$ increases which makes roll control more difficult. This suggests that higher values of $\lambda_s$ would measure the overall level of difficulty. Two summary measures of $\lambda_s$ were calculated for each run: the maximum value attained ($\lambda_{max}$) and the mean ($\lambda_{mean}$). While pilot comments indicated that actuation of the roll loop made control more difficult and generally less acceptable, these measures were found to have no analytical power for system handling quality evaluation.

The shape of the logistic curve is determined by the four parameters $B_1$-$B_4$. To better interpret the estimated control laws, the parameters $B_1$-$B_4$ are rescaled into new parameters $P_1$-$P_4$. Each of the parameters $P_1$-$P_4$ relates to a specific geometric property of the curve. $P_1$ is the control value given by the logistic curve for zero tracking error; $P_2$ is the tracking error at the center of the logistic curve; $P_3$ is the slope of the central linear portion of the logistic curve; $P_4$ is the tracking error interval over which the logistic control law is linear. These physical characteristics of the parameters $P_1$-$P_4$ are presented in Figure 13. For illustration of the variety of shapes taken on by logistic curves, four different logistic curves and their associated parameters are presented in Figure 14.

**Figure 12.** Examples of the Variability of Instability Parameter $\lambda_s$

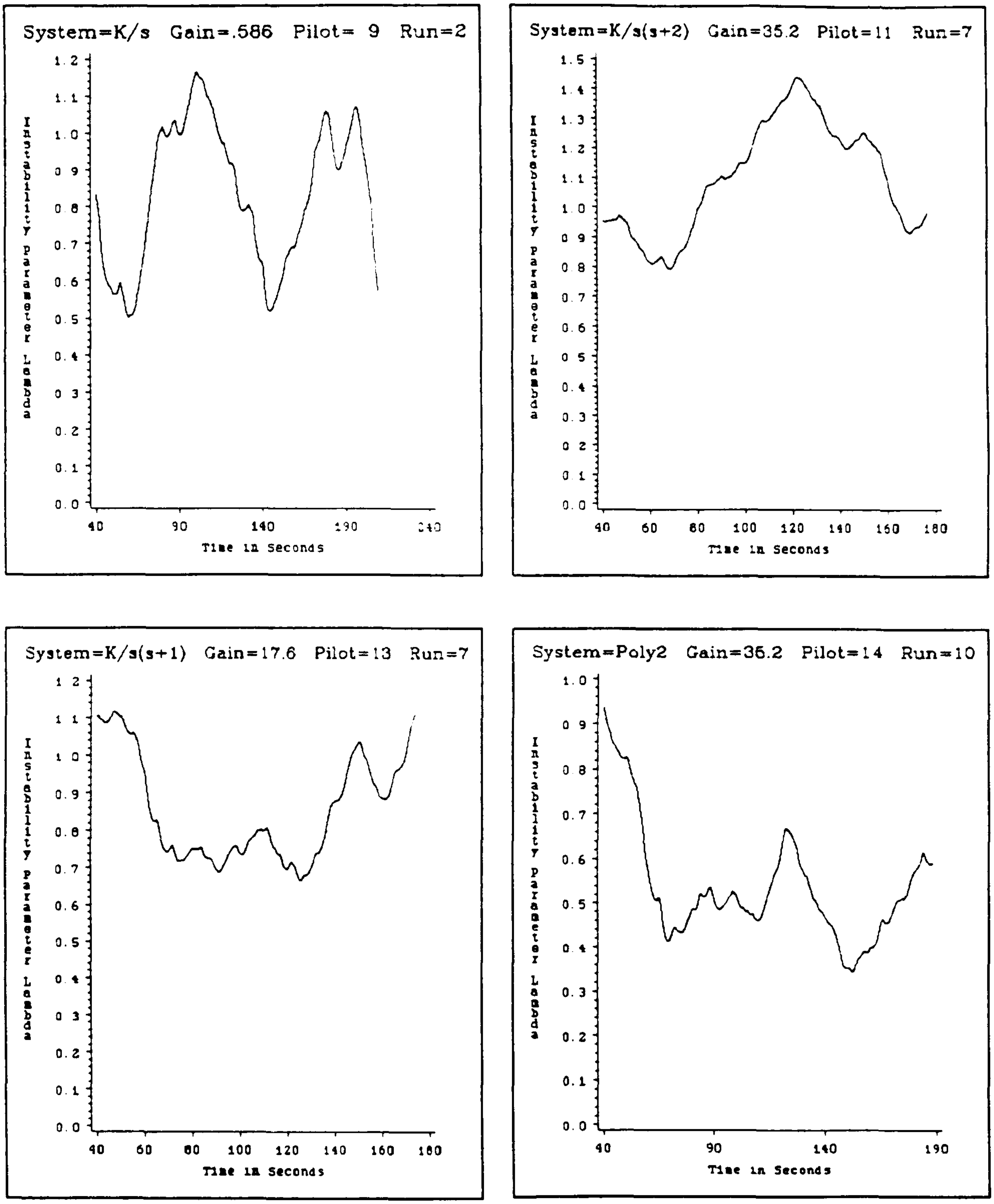


**Figure 13.** Physical Characteristics of $P_1$, $P_2$, $P_3$ and $P_4$

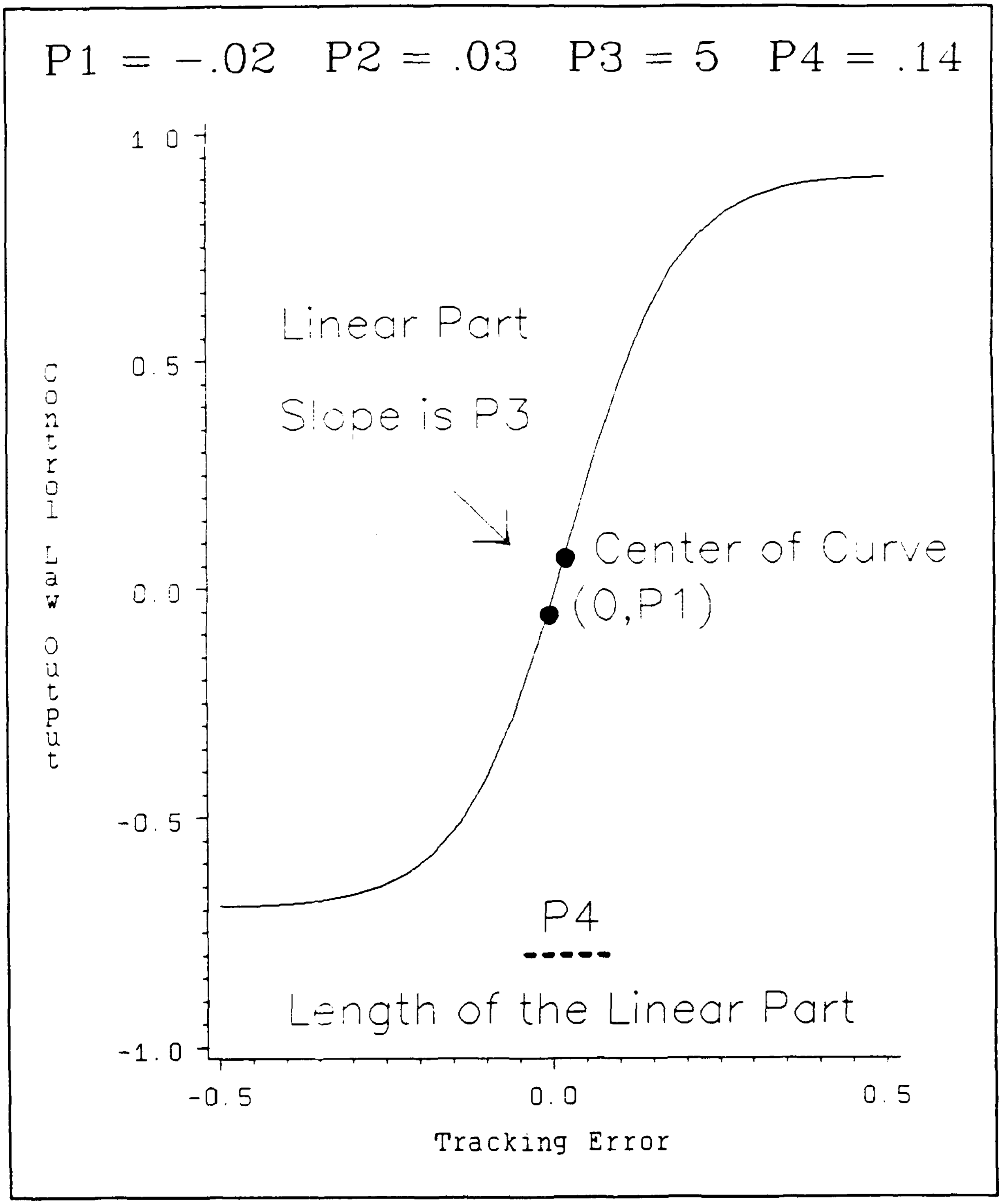
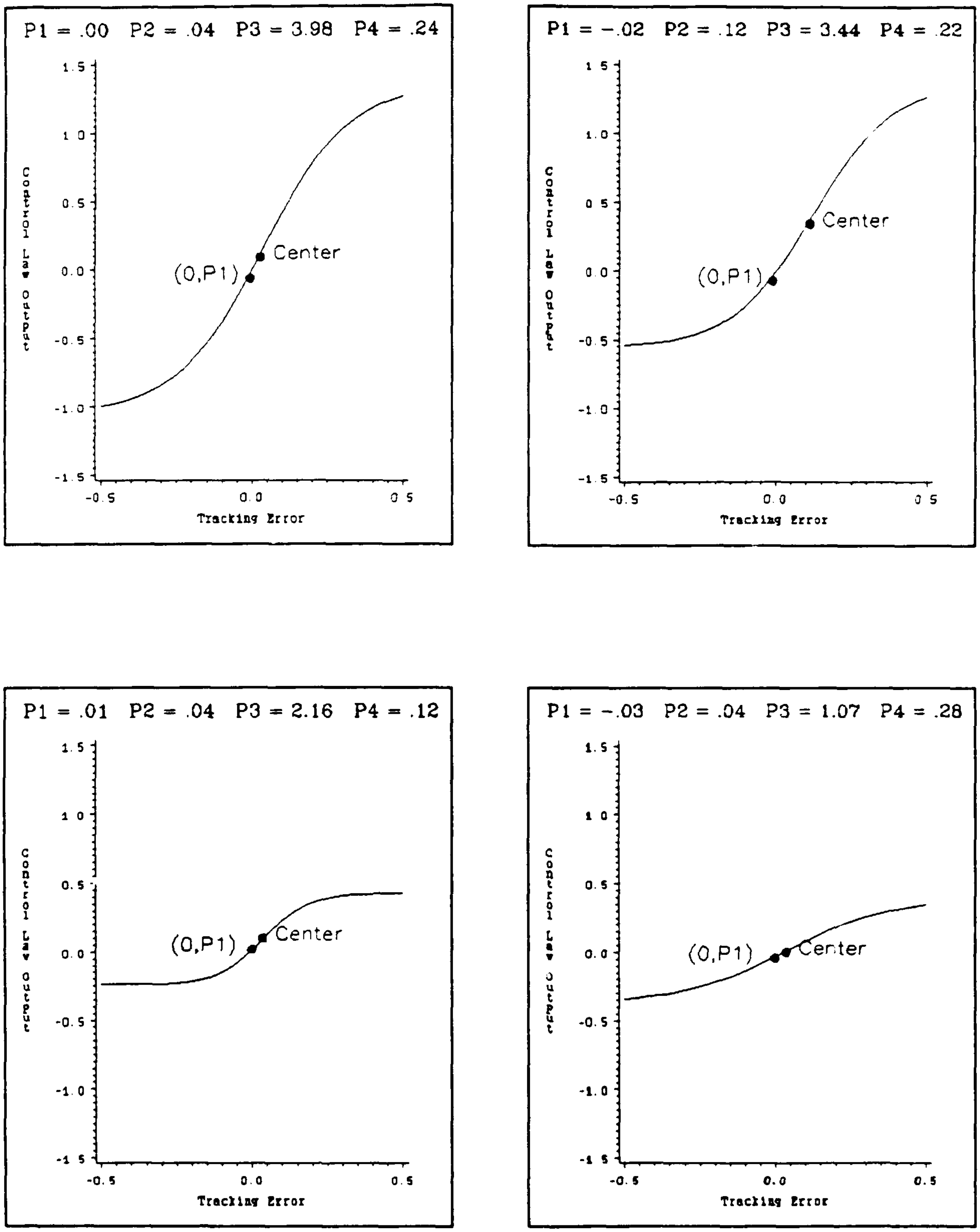


**Figure 14.** Logistic Curves for Various Values of $P_1$, $P_2$, $P_3$ and $P_4$



## E. Scaling Associated with Screen Vertical Display Scale and Bandwidth of $\theta_c$

Both the screen vertical display scale and the bandwidth of $\theta_c$ affect the magnitude of the pitch tracking target. Thus, the pilot must vary the scale of his pitch input to keep up with the displayed tracking target driven by $\theta_c$. That is, the scale of $\delta_e$ and $\theta_e$ are also affected.

In the next section, the values of the primary and secondary performance measures are compared over all the runs grouped according to system. When the runs differ on screen vertical display scale and/or bandwidth of $\theta_c$, this comparison is adversely affected by this scale change in $\theta_c$, $\delta_e$, and $\theta_e$. To compensate for this scaling, scaled values of the "affected" performance measures are used for comparison. The "affected" measures are RMS, EMS, MEAN, MEAN_RUN, $P_1$, $P_2$, and $P_4$ which are divided by a scale factor depending upon the screen vertical display scale and the bandwidth selected for the run.

The scale factor is simply a product of the scaling associated with the screen vertical display scale and that associated with the bandwidth. The scaling associated with the screen vertical display is straightforward since $\theta_c$ is multiplied by 0.90 for "normal" screen vertical display and 1.56 for "wide". To define the scaling associated with bandwidth, the root mean square of $\theta_c$ over 100 seconds is used. ($\theta_c$ is periodic over 100 seconds.) For "low" bandwidth, it is .176635 while for "high" bandwidth, it is .20199. Thus, the scale factors used are:

| | | Bandwidth of $\theta_c$ | |
|---|---|---|---|
| | | Low | High |
| Screen Vertical Display Scale | Normal | 0.90 × .176635 | 0.90 × .20199 |
| | Wide | 1.56 × .176635 | 1.56 × .20199 |

## F. Summary of Pilot Performance Measures

In summary, the performance measures analyzed are as follows:

Primary Measures

Error measures of the statistical estimation of the pilot control law (RMS, EMS and RSQ).

Mean absolute tracking error (actual run, MEAN_RUN; and simulated run, MEAN).

Secondary Measures

Secondary task (roll) stability parameter ($\lambda_{max}$ and $\lambda_{mean}$).

Control Law Geometric Characteristics ($P_1$-$P_4$).

In the following paragraphs, these performance measures (some appropriately scaled) will be analyzed and related to both handling qualities and the subjective Cooper-Harper rating.

# SECTION IV
# RELATIONSHIP BETWEEN PERFORMANCE MEASURES AND SYSTEMS

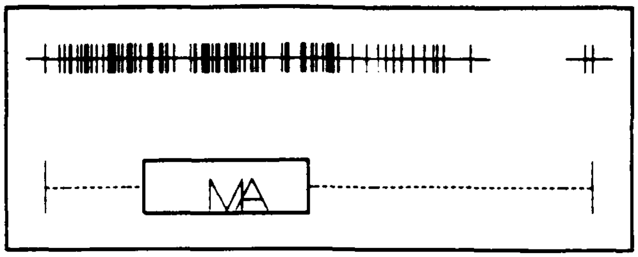
## A. Overview

The previous section describes the definition and calculation of derived variables that condense the measured experimental data on pilot response into measures of pilot performance that are potentially useful for the evaluation of handling qualities. This section presents an analysis of these performance measures focusing on the ability of the performance measures to effectively differentiate the seven systems used in the experiments. The way that the performance measures differentiate between systems is a basis for explaining the real differences between system handling qualities from a pilot control standpoint. Presented first, for each of the performance measures, is a one-dimensional view of that measure's ability to differentiate the systems. This presentation is supported through graphics which utilize box plots. Then follows a statistical analysis of the ability of the measures jointly to differentiate the systems. Statistically, the general term for such an analysis is discrimination analysis. Several statistical techniques for discrimination analysis will be utilized in the data analysis to show different aspects of the relationship between the performance measures and handling qualities. The performance measures will be ranked by their relative importance in the discrimination analysis. The specific role of each performance measure in the finer levels of differentiating systems is discussed. The conclusion is that the performance measures are effective analysis variables for the differentiation among systems and thus for evaluating handling qualities.

## B. General Characteristics of the Data

Seven different input/output transfer functions are used to define the systems flown in the experiments. These systems are defined earlier in this report. For each analysis performance measure and system, the values for that measure are summarized by a box plot. A box plot is a statistical graphic which is used in much of the analysis in this report. This graphic presents a quick visual summary of percentile statistics for data. A box covers the interquartile range of data (from the 25th through the 75th percentiles) and whiskers extend out to the minimum data value on the left and the maximum data value on the right; "M" marks the location of the median (50th percentile) of the set of data and "A" the arithmetic average of the data. Figure 15 demonstrates the association between this summary and a set of data.

**Figure 15.** Box Plot Example for a Set of Data



The graphics presented in Figures 16 through 26 summarize the variation in the performance measures across these seven systems using box plots. In the graphic summary for a particular performance measure, the box plots are arrayed in the order of increasing median values to visually display how that performance measure "discriminates" between the seven systems.

## B.1. Primary Performance Measures

Figures 16 through 20 present the variation across systems of the primary performance measures. These figures indicate that each of the performance measures provide a partial discrimination between systems.

Figure 16 summarizes the way that the regression mean square (RMS) differentiates between systems. RMS is a measure of the variability in the pitch control input which is prescribed by the estimated control law. If the estimated control law is flat, then RMS will be low; whereas a steep estimated control law or a control law with significant curvature will have a large RMS. RMS basically separates the systems into three groups:

K/s(s+1) and Poly1 have the lowest RMS values.

K/s(s+4) and K/s(s+2) have intermediate RMS values.

K/s, K/$s^2$ and Poly2 have the highest RMS values.

**Figure 16.** Regression Mean Square (Scaled)
Box Plot Summary by Systems

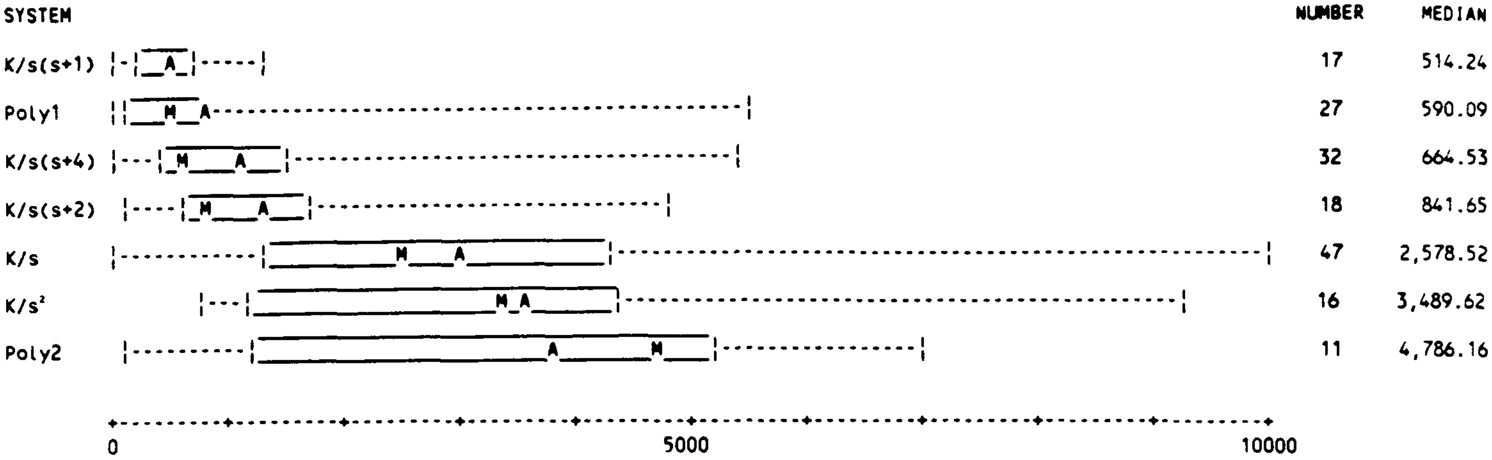


Figure 17 presents the way that the error mean square (EMS) differentiates between systems. EMS is a measure of the variability of the actual run data around the estimated control law curve. It is a measure of how well the pilot was able to maintain his ideal control law; that is, it is a measure of the difference between predicted and actual pilot inputs. EMS separates the systems into five groups:

K/s(s+1) has the lowest EMS values.

K/s(s+4), K/s(s+2) and Poly1 have similar higher EMS values.

K/s has EMS values slightly higher typically than the previous group.

Poly2 has EMS values much higher than the previous group.

K/s$^2$ has the highest EMS values.

**Figure 17.** Error Mean Square (Scaled)
Box Plot Summary by System

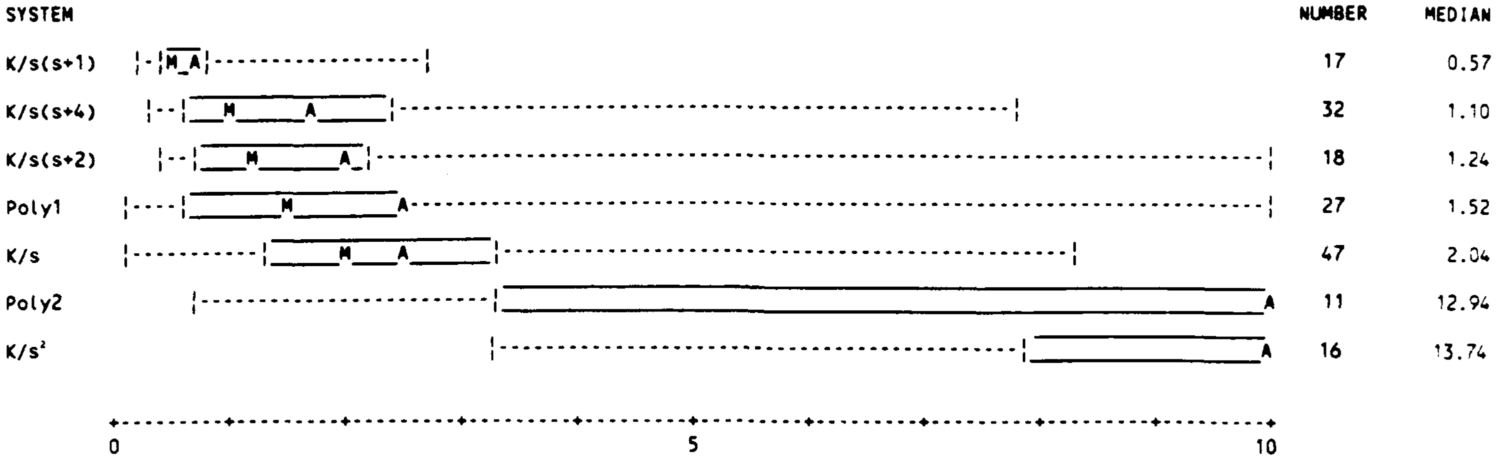


Figure 18 presents the way that the R-square value (RSQ) differentiates between systems. RSQ is the ratio of the regression sum of squares and the total sum of squares for the data and thus can take on values only between 0 and 1. The regression sum of squares is four times RMS, since the regression mode has four degrees of freedom. To calculate the total sum of squares for the data, for each data point take the squared deviation between the actual pilot input and the run average pilot input, and sum these squared deviations over all data points in the run. RSQ is a measure of how much of the total sum of squares comes from the regression sum of squares. Therefore, RSQ is a measure of how well the estimated control law "fits" the data with RSQ = 1 being a perfect fit. RSQ hierarchically separates the systems into five groups:

Poly2 and K/s$^2$ have the lowest RSQ values.

Poly1 has slightly higher RSQ values.

K/s(s+4) has significantly higher RSQ.

K/s(s+2) and K/s(s+1) RSQ values are somewhat higher.

K/s has somewhat higher RSQ values.

**Figure 18.** R-Square
Box Plot Summary by Systems

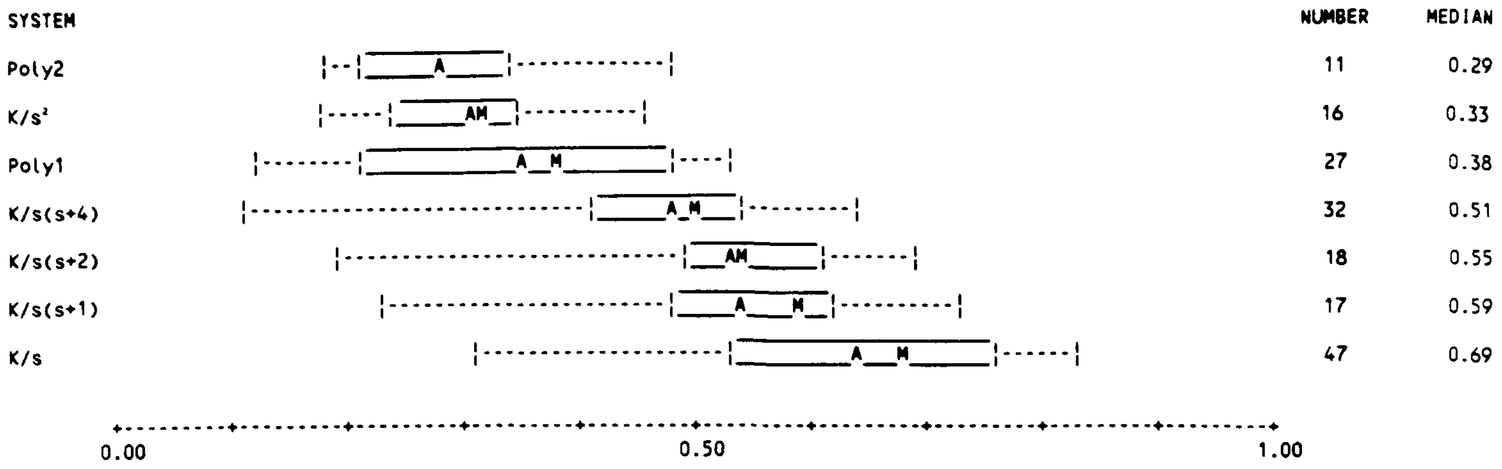


Figure 19 presents the way that the mean absolute tracking error from the actual run (MEAN_RUN) differentiates between systems. MEAN_RUN clearly separates K/s$^2$ as having much higher tracking errors than the other six systems. The remaining six systems all have low and similar tracking error patterns.

**Figure 19.** Mean Absolute Tracking Error From Data
Box Plot Summary by Systems

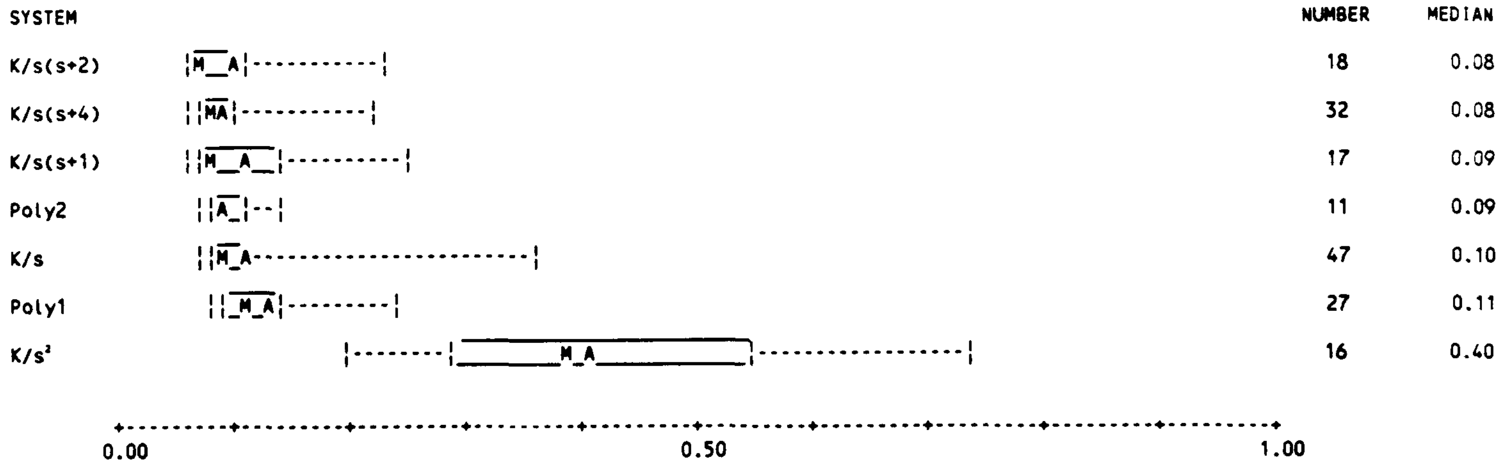


Figure 20 presents the way the mean absolute tracking error from the simulation run (MEAN) (using the estimated control law to calculate pilot input in a closed loop) differentiates between systems. MEAN cleanly separates the systems into three groups:

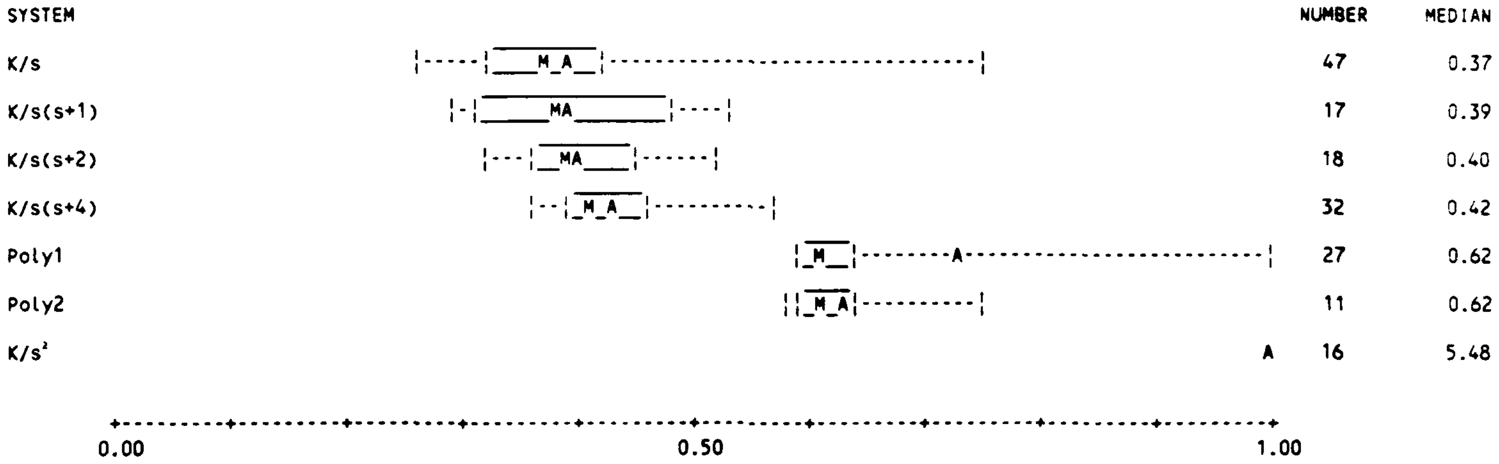
K/s, K/s(s+1), K/s(s+2) and K/s(s+4) have the lowest MEAN values.

Poly1 and Poly2 have higher MEAN values

K/$s^2$ has very high MEAN values.

There is a slight ordering of the systems in the first group, but the systems are not separated. Note that there is better discrimination between systems by MEAN than there is by MEAN_RUN, the mean absolute tracking error from the data run. Partial explanation for the poorer discrimination power in MEAN_RUN comes from the fact that as an error measure MEAN_RUN includes errors associated with pilot lag whereas MEAN does not.

**Figure 20.** Mean Absolute Tracking Error From Simulation
Box Plot Summary by Systems



## B.2. Secondary Performance Measures

Figures 21 through 26 present the variation across systems of the secondary performance measures.

Figure 21 presents the way that the maximum value of $\lambda_s$ ($\lambda_{max}$) recorded during the run differentiates between systems for runs involving the secondary loading task. $\lambda_s$ is the instability parameter for the roll control in the tracking task. The roll component of the tracking task is in the system to maintain a constant level of difficulty for the overall tracking task. The pattern of the box plots in Figure 21 seems to indicate that the less complex systems have somewhat higher $\lambda_{max}$ values; this probably reflects a higher tolerance by the pilot for instability in the roll component of the less complex systems. The exception is with $K/s^2$. $\lambda_s$ is modulated by the difference between the current tracking error and an accumulated average tracking error. The exceptionally large tracking errors experienced in the $K/s^2$ runs prevent meaningful comparison of $\lambda_s$ values in $K/s^2$ runs to $\lambda_s$ values in the runs using the other systems.

**Figure 21.** Maximum of $\lambda_s$
Box Plot Summary by Systems

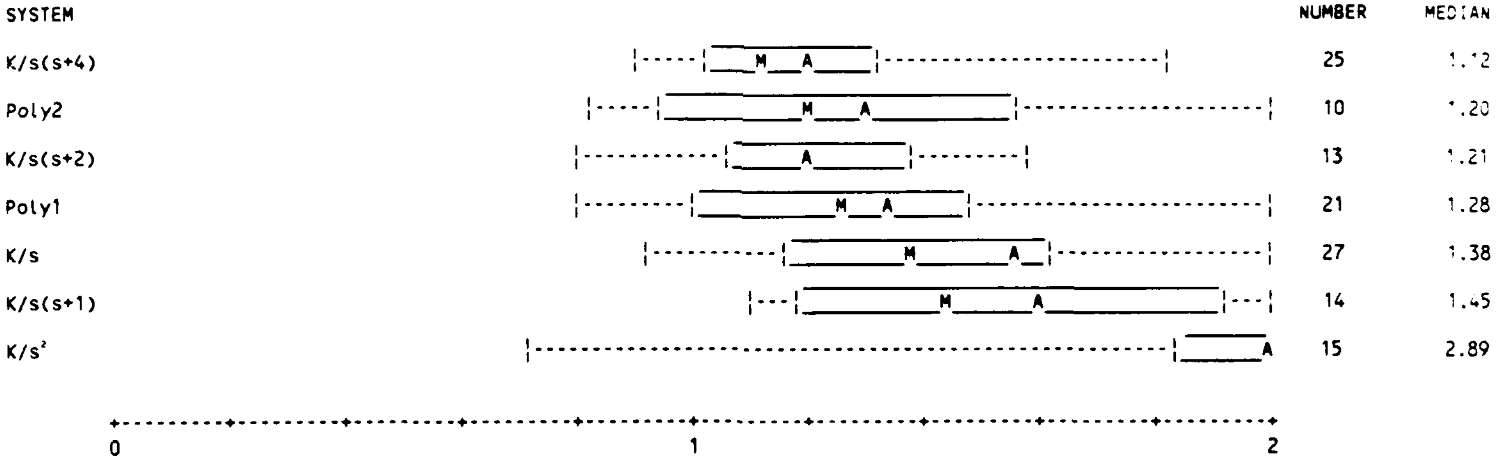
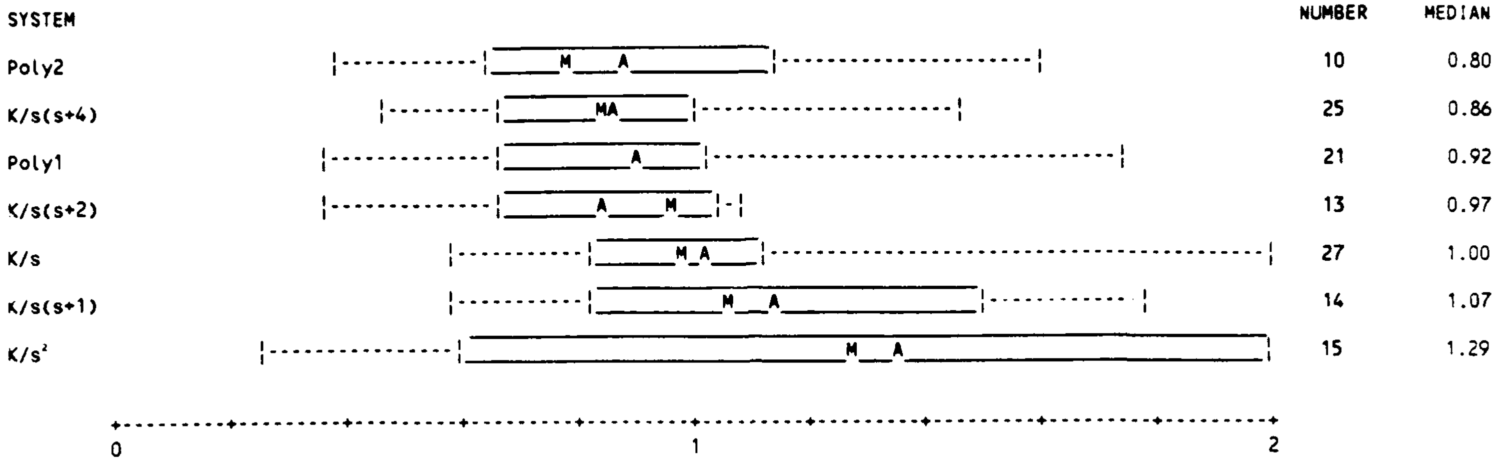


Figure 22 presents the way that the average value of $\lambda_s$, $\lambda_{mean}$, recorded during the run differentiates between systems for runs involving the secondary loading task. The information in $\lambda_{mean}$ is virtually identical to the information in $\lambda_{max}$ presented in Figure 21 above.

**Figure 22.** Mean of $\lambda_s$
Box Plot Summary by Systems



Figures 23 through 26 present the way that the geometric parameters $P_1$-$P_4$ of the estimated pilot control law differentiate between systems. These parameters are defined above in Section III.D on Secondary Measures of Pilot Performance.

Figure 23 presents the way that parameter $P_1$ differentiates between systems. $P_1$ is the input control value given by the estimated control law for zero tracking error. For all of the systems, except $K/s^2$, this control value is typically negative.

**Figure 23.** Control Law Parameter $P_1$
Box Plot Summary by Systems

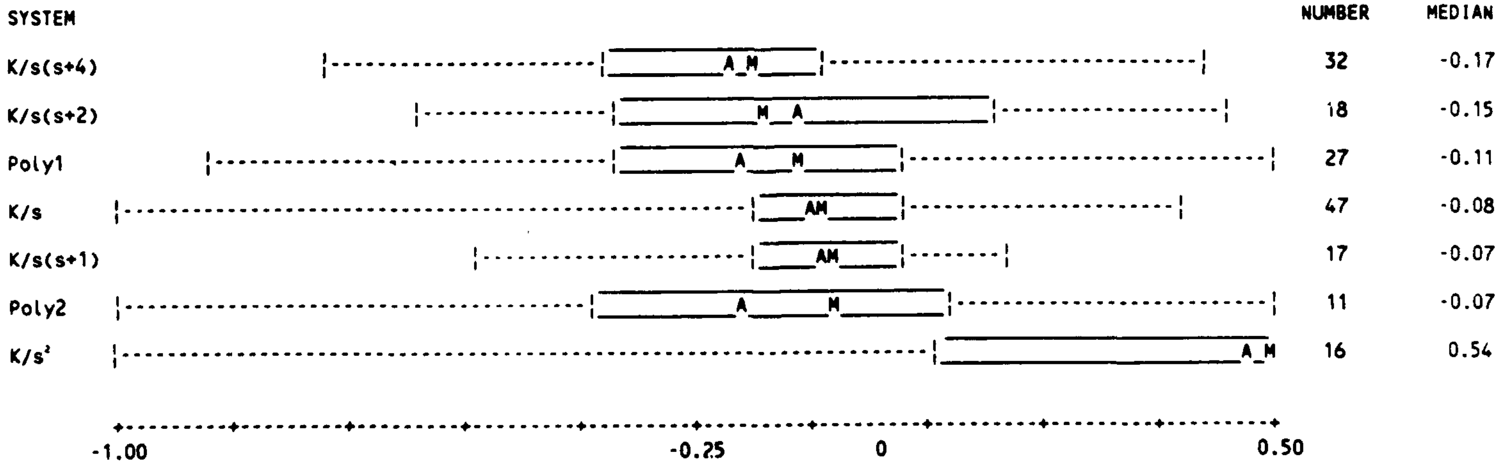
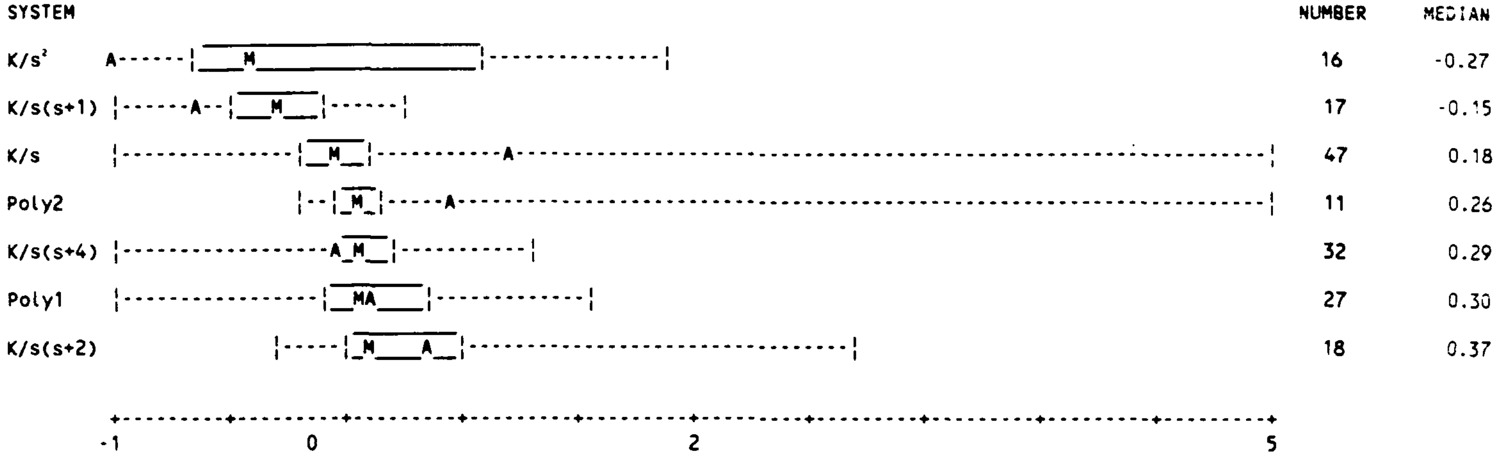


Figure 24 presents the way that parameter $P_2$ differentiates between systems. $P_2$ is the tracking error at the center of the estimated control law curve. $P_2$ is typically negative for two of the systems and positive for the other five systems. $P_2$ separates the systems into two groups:

> $K/s^2$ and $K/s(s+1)$ have typically negative values of $P_2$.
>
> $K/s$, Poly2, $K/s(s+4)$, Poly1 and $K/s(s+2)$ have typically positive values of $P_2$.

**Figure 24.** Control Law Parameter $P_2$
Box Plot Summary by Systems



| SYSTEM | NUMBER | MEDIAN |
|---|---|---|
| K/s² | 16 | -0.27 |
| K/s(s+1) | 17 | -0.15 |
| K/s | 47 | 0.18 |
| Poly2 | 11 | 0.26 |
| K/s(s+4) | 32 | 0.29 |
| Poly1 | 27 | 0.30 |
| K/s(s+2) | 18 | 0.37 |

Figure 25 presents the way that parameter $P_3$ differentiates between systems. $P_3$ is the slope of the central linear portion of the estimated control law curve. Larger $P_3$ values will lead to larger variability in the input control values given by the estimated control law curve; this in turn will lead to large regression mean square (RMS) values for the control law model. Large curvature, as well, can lead to large RMS values. Note that the differentiation between systems by RMS as presented in Figure 16 is almost identical with the differentiation by $P_3$ presented in Figure 25. The only real difference is with the $K/s^2$ system. It is, in fact, the high curvature, rather than the linear slope, in the $K/s^2$ estimated control laws which causes this system to have high RMS values. $P_3$ separates the systems into three groups:

> Poly1, $K/s(s+1)$ and $K/s^2$ have the lowest $P_3$ values.
>
> $K/s(s+4)$ and $K/s(s+2)$ have intermediate $P_3$ values.
>
> $K/s$ and Poly2 have the highest $P_3$ values.

**Figure 25.** Control Law Parameter $P_3$
Box Plot Summary by Systems

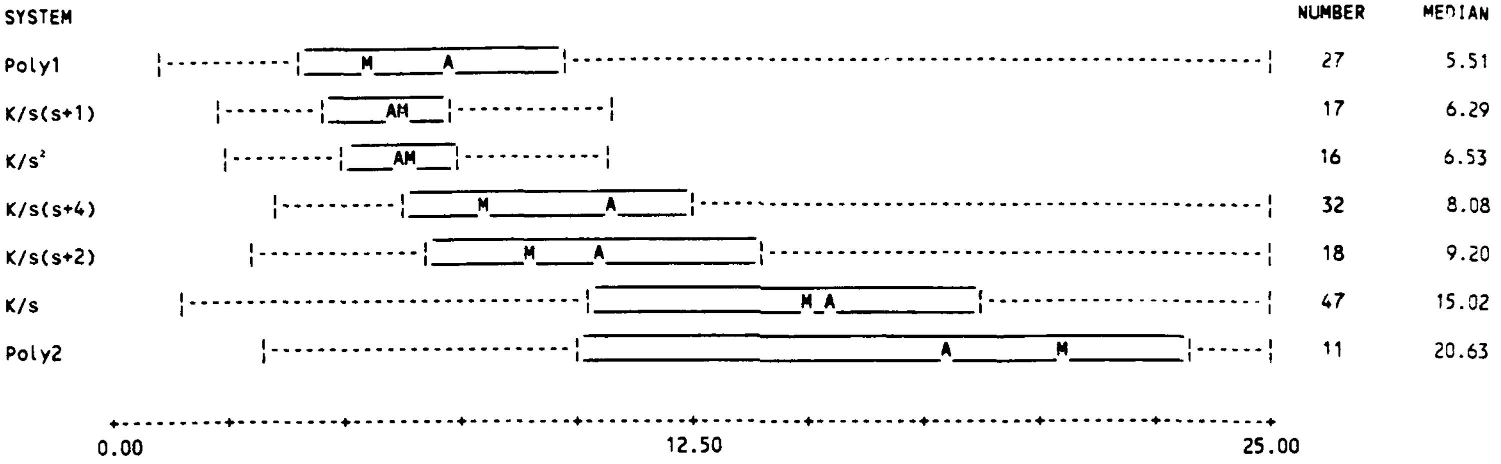
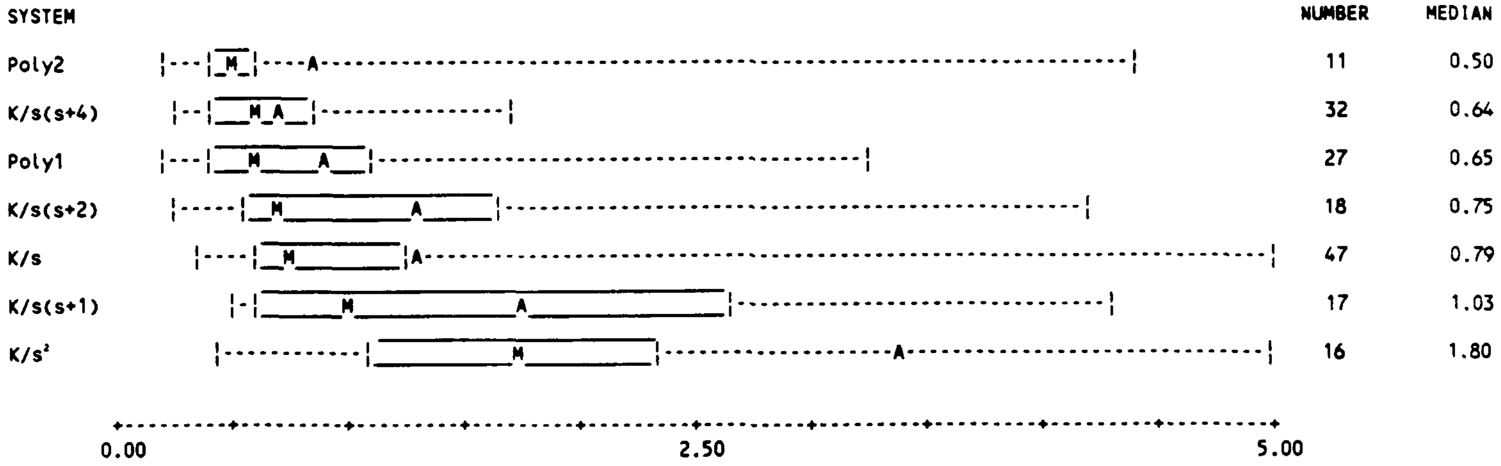


Figure 26 presents the way that parameter $P_4$ differentiates between systems. $P_4$ is the tracking error interval over which the estimated control law curve is linear. The smaller $P_4$ is, the quicker the "saturation" of the estimated control law curve sets it. Saturation means that the control law input becomes less, in absolute value, than it would if the linear part of the curve were extended. Simply put, saturation means an easing off of control inputs for larger tracking errors. Figure 26 presents a hierarchy of the systems by parameter $P_4$ with no adjacent systems in the hierarchy differing by very much with respect to $P_4$.

**Figure 26.** Control Law Parameter $P_4$
Box Plot Summary by Systems



## C. Joint Analysis of Performance Measures using Discrimination Analysis

Now that the performance measures have been introduced and discussed singly for their ability to differentiate systems, the following analysis considers the ability of the measures jointly to provide information about the systems. The statistical tools that are utilized are called discrimination analysis; several different algorithms for discrimination analysis are considered. The performance measures are ranked by their relative importance in the discrimination analysis and the specific role of each performance measure in the finer levels of differentiating systems is discussed. The conclusion of the analysis is that the performance measures are effective analysis variables for the differentiation among systems and thus among handling qualities.

### C.1. Discrimination Algorithms

A general discrimination problem concerns the classification of individuals into groups using classification variables. The classification variables can be qualitative as well as quantitative although many discrimination algorithms are based on a Gaussian distributional assumption for the classification variables within each group which restricts the classification variables to being quantitative. Classification rules based on the classification variables are estimated by minimizing some error measure associated with "misclassifying" individuals. These classification rules can then be used to:

> Characterize the groups according to how they differ on values of the classification variables.
>
> Determine the classification variables with the "greatest" discrimination power to differentiate among groups.

In this study, the groups are the different experimental systems and the classification variables are the performance measures defined previously. The classification rules help to determine how a pilot's selected control law and accuracy vary depending upon which system is being flown. This information is used to characterize a system and thus evaluate handling qualities via the reaction of a pilot to it.

Three discrimination algorithms are applied to the data. Two of the algorithms are available in SAS (1985) (Procedures CANDISC and STEPDISC); the third is a classification tree algorithm implemented in a special-purpose software (CART). Appendix D contains a brief description of these algorithms. Each of these algorithms is founded upon a probabilistic model which is assumed to underlie the mechanism generating the data. Some algorithms are known to perform well even when their associated model is violated. Each of these techniques is applied to our data in order to compare and contrast the information which they provide concerning the relationship between performance measures and handling qualities.

### C.2. Application of Statistical Discrimination Algorithms - STEPDISC and CANDISC

The CANDISC procedure is used to provide a linear ranking of the systems in the spirit of a loss function index as investigated by McDonnell (1968) and Baldwin and Gantz (1983, 1984). CANDISC is based on finding linear combinations of performance measures which are most correlated with systems. Thus, CANDISC provides some information about a linear ordering of the systems based on values of a linear function of performance measures.

Since CANDISC does not explicitly rank variables by importance, an initial investigation of which performance measures to use in the linear combinations is done using STEPDISC, another SAS Procedure.

STEPDISC ranks variables according to the statistical partial correlation for predicting the value of a variable from a model for group classification; the model controls for the effects of the other variables already in the model. The variable set

$$\text{MEAN\_RUN, MEAN, } P_1\text{-}P_4\text{, EMS, RMS, and RSQ}$$

was used in the application of STEPDISC. STEPDISC using the STEPWISE option with significance level 0.15 for a variable to enter the model and to stay produced the following ranking:

$$\text{MEAN} \Rightarrow \text{RSQ} \Rightarrow \text{MEAN\_RUN} \Rightarrow \text{EMS} \Rightarrow P_3 \Rightarrow P_1 \Rightarrow P_2 \Rightarrow \text{RMS}$$

$P_4$ is not selected under these enter and stay levels. (See Appendix D.) Table 4 shows the squared partial correlations themselves as well as the squared correlations. The squared correlation is an indication of the performance measure's isolated ability to discriminate between systems, whereas the squared partial correlation is an indication of the additional discrimination in the model when the performance measure is added.

**Table 4.** Discrimination Power of Performance Measures
Measured by Squared Correlation and Squared Partial Correlation with Systems

| Performance Measure | Sq. Partial Correlation | Squared Correlation |
|---|---|---|
| MEAN | .826 | .826 |
| RSQ | .433 | .498 |
| MEAN_RUN | .394 | .677 |
| EMS | .333 | .527 |
| $P_3$ | .124 | .255 |
| $P_1$ | .102 | .195 |
| $P_2$ | .075 | .071 |
| RMS | .060 | .321 |
| $P_4$ | .058 | .106 |

Table 4 suggests that MEAN, RSQ, MEAN_RUN, and EMS are the "main" performance measures useful in discriminating systems. The high correlation of MEAN with system indicates that systems vary significantly as to how well the pilot can perform the given task using the handling qualities embodied in the system. The table also points out the relationship of MEAN and RSQ to the other variables in terms of discriminating power; once MEAN and RSQ are in the model, the squared correlation between system and other performance measures is greatly reduced as seen by comparing the squared partial correlation to the squared correlation. This means that much of the information about handling qualities contained in the other performance measures is duplicated in MEAN and RSQ.

After STEPDISC provided this breakdown of the importance of the potential classification variables, CANDISC was run on five different subsets of the performance measures:

| | |
|---|---|
| Set A: | MEAN, RSQ, EMS, MEAN_RUN, RMS, $P_1$-$P_4$ |
| Set B: | MEAN, RSQ, EMS, MEAN_RUN, RMS |
| Set C: | MEAN, RSQ, EMS, MEAN_RUN |
| Set D: | MEAN, EMS |
| Set E: | MEAN, RSQ |

Different subsets of performance measures are considered to further determine the relative importance of different variables in providing information about handling qualities. In particular, the reason for considering Sets D & E, each with two performance measures, is to generate a linear combination on similar variables to those suggested in McDonnell (1968) and Baldwin and Gantz (1983, 1984). Baldwin and Gantz considered a loss function based upon a linear combination of the pilot error in maintaining his desired control law (EMS) and the error of the pilot's control as summarized in mean absolute error of $\theta_e$ (MEAN or MEAN_RUN).

Due to the variety of the systems used, it is potentially important to consider the pilot error in maintaining his desired control law in a standardized form in terms of RSQ since

$$\text{RSQ} \equiv 1 - (\text{error sum of squares})/(\text{total sum of squares}).$$

RSQ is a measure of the correlation between the pilot's "ideal" control law and his measured output and is thus also a measure of how well the pilot maintained his ideal control law.

The results of the CANDISC runs are summarized in Table 5. The first part of the table presents the coefficients for each performance measure in the linear combination of performance measures which is used to linearly order systems. The linear combination of performance measures which linearly orders systems is called the canonical variable. The coefficients defining the linear combination of performance measures are called canonical coefficients. The second part of the table presents the total statistical correlation between each performance measure and the canonical variable. The total correlation tells how the performance measure is related to the canonical variable. Correlations close to one or negative one are strong correlations.

**Table 5.** Summary of Canonical Discrimination Analysis
Various Sets of Performance Measures

Canonical Coefficients

| | MEAN | RSQ | MEAN_RUN | EMS | $P_3$ | $P_1$ | $P_2$ | RMS | $P_4$ |
|---|---|---|---|---|---|---|---|---|---|
| Set A | 1.75 | -.17 | 1.35 | .43 | -.14 | .33 | -.22 | .06 | -.11 |
| Set B | 1.83 | -.09 | 1.24 | .45 | | | | -.04 | |
| Set C | 1.83 | -.11 | 1.24 | .42 | | | | | |
| Set D | 2.24 | | | .21 | | | | | |
| Set E | 2.28 | -.23 | | | | | | | |

Total Correlation Between Performance Measures and Canonical Variables

| | MEAN | RSQ | MEAN_RUN | EMS | $P_3$ | $P_1$ | $P_2$ | RMS | $P_4$ |
|---|---|---|---|---|---|---|---|---|---|
| Set A | .95 | -.40 | .85 | .61 | -.22 | .43 | -.22 | .27 | .30 |
| Set B | .96 | -.40 | .86 | .62 | | | | .28 | |
| Set C | .96 | -.40 | .86 | .62 | | | | | |
| Set D | .99 | | | .64 | | | | | |
| Set E | .99 | -.43 | | | | | | | |

| | Set A | Set B | Set C | Set D | Set E |
|---|---|---|---|---|---|
| Sq. Canonical Corr. (CANRSQ) | .914 | .901 | .901 | .828 | .829 |
| CANRSQ/(1-CANRSQ) | 10.620 | 9.080 | 9.070 | 4.820 | 4.850 |

CANRSQ is the squared multiple correlation between the system groups and the canonical variable.
CANRSQ/(1-CANRSQ) can be interpreted as the ratio of between class variation to within class variation for canonical variables.

The correlations between performance measures and the canonical variable show that the canonical variable is dominated by MEAN. The next most correlated variable is MEAN_RUN. However, as seen in Table 4, MEAN_RUN loses much of its discrimination power once MEAN is in the model. Thus, the high correlation with MEAN_RUN is due mostly to the high correlation between MEAN and MEAN_RUN. Comparing CANRSQ/(1-CANRSQ) over different variable sets shows that little is lost by eliminating $P_1$-$P_4$ (Set B) and RMS (Set C). However, the ability to differentiate between systems is diminished by further removing MEAN_RUN, EMS, or RSQ (Sets D and E). This indicates that the majority of the information about handling qualities contained in the performance measures is contained in MEAN, RSQ, MEAN_RUN, and EMS.

One of the options of CANDISC is a pairwise test of the difference between systems based on the Mahalanobis distance between systems. (The Mahalanobis distance between systems is the Euclidean distance between the within-system means of the performance measures weighted by the within-system covariances.) Table 6 shows the systems which are not significantly different at a significance level of 0.10 which reveals that different performance measures are important in discriminating between different systems. For example, $P_1$-$P_4$ and RMS have little discriminating power in the presence of the other variables in Set C. EMS is important in discriminating between Poly1 and Poly2 as seen by comparing Set E to Set C while RSQ is important in discriminating K/s from K/s(s+a), a=1,2,4, runs as seen by comparing Set D to Set C.

**Table 6.** Significantly Different Systems Based on Mahalanobis Distance Using Various Performance Measure Sets. Underlined Systems Are Not Significantly Different at a 0.10 Significance Level.

<table>
<tr><td>Set A</td><td colspan="3">K/s(s+1) K/s(s+2) K/s(s+4)</td><td>K/s</td><td>Poly1</td><td>Poly2</td><td>$K/s^2$</td></tr>
<tr><td>Set B</td><td colspan="3">K/s(s+1) K/s(s+2) K/s(s+4)</td><td>K/s</td><td>Poly1</td><td>Poly2</td><td>$K/s^2$</td></tr>
<tr><td>Set C</td><td colspan="3">K/s(s+1) K/s(s+2) K/s(s+4)</td><td>K/s</td><td>Poly1</td><td>Poly2</td><td>$K/s^2$</td></tr>
<tr><td>Set D</td><td colspan="5">K/s(s+1) K/s(s+2) K/s(s+4) K/s Poly1</td><td>Poly2</td><td>$K/s^2$</td></tr>
<tr><td>Set E</td><td colspan="3">K/s(s+1) K/s(s+2) K/s(s+4)</td><td>K/s</td><td colspan="2">Poly1 Poly2</td><td>$K/s^2$</td></tr>
</table>

Table 6 also shows that with the performance measures MEAN, RSQ, MEAN_RUN, and EMS, one can effectively differentiate between all of the systems except for those of the form K/s(s+a), a=1,2,4. Thus, the performance measures can be used to group like systems, (i.e., like handling qualities) up to a degree; some minor differences between systems cannot be detected by the techniques used so far.

It is important to note that besides grouping the systems, the results of CANDISC can also be used to linearly rank the systems much like a Cooper-Harper rating. The medians of the canonical variable within each system provides this linear ordering of the systems. Table 7 shows the system medians for the canonical variables from the different performance measure sets. Based on the linear ordering supplied by the canonical variable, regardless of the variable set used, the linear combination can separate the simpler systems K/s, K/s(s+a), the polynomial systems, and $K/s^2$ ranking them in this order from best to worst. The different variable subsets, however, have varied success in separating within the systems K/s and K/s(s+a), a=1,2,4.

**Table 7.** System Medians for the Canonical Variables
Numbers in Parentheses Are the Ranking of the Systems Based on Increasing Values of the Medians. The Last Column Shows the Median Cooper-Harper Rating within System.

| | Set A | Set B | Set C | Set D | Set E | Cooper-Harper Rating |
|---|---|---|---|---|---|---|
| K/s(s+1) | -1.44(3) | -1.48(1) | -1.49(1) | -.89(1) | -.82(3) | 5.00 |
| K/s(s+2) | -1.39(4) | -1.47(2) | -1.48(2) | -.85(2) | -.86(2) | 4.25 |
| K/s(s+4) | -1.52(1) | -1.37(3) | -1.37(3) | -.81(4) | -.78(4) | 4.00 |
| K/s | -1.47(2) | -1.33(4) | -1.28(4) | -.84(3) | -1.02(1) | 5.00 |
| Poly1 | -.59(5) | -.69(5) | -.69(5) | -.58(5) | -.26(5) | 6.00 |
| Poly2 | -.00(6) | .01(6) | .04(6) | -.22(6) | -.19(6) | 6.00 |
| $K/s^2$ | 9.37(7) | 8.44(7) | 8.45(7) | 6.07(7) | 6.04(7) | 9.00 |

Note the canonical variable clearly distinguishes between Poly1 and Poly2 for which the Cooper-Harper rating is the same. On the other hand, the Cooper-Harper rating suggests that the K/s(s+a) and K/s systems can be ordered as K/s(s+4), K/s(s+2), with K/s(s+1) and K/s comparable. This is an example of the complementary information which can be obtained via combining the objective and subjective measures of pilot performance.

## C.3. Application of Statistical Discrimination Algorithms - CART

The third statistical discrimination algorithm applied to the test data is CART (Classification and Regression Trees). CART is used to classify the experimental runs into the seven systems based on the performance measures defined above. Details about CART can be found in Appendix D.

CART has two advantages when compared to STEPDISC and CANDISC. First, CART "automatically" allows for heterogeneity with regards to different performance measures being important in discriminating between different subsets of systems. For the other two algorithms, multiple runs of statistical analysis have to be performed with different subsets of performance measures and systems to "discover" which variables are important in discriminating between which systems. Second, the tree-like structure of the resulting CART discrimination rule makes the CART models easy to interpret. Additionally, similar to STEPDISC, CART provides a rank ordering of the importance of variables in discriminating systems, a feature which is useful in variable selection.

One "defect" of CART is that it does not provide a linear ranking of the systems in the spirit of the CANDISC analysis earlier in this section. CART only provides information about which variables are important for discriminating between systems in a many-variable environment. The following table summarizes the performance measures presented to CART and CART's assessment of the relative value of each performance measure for discriminating between the systems.

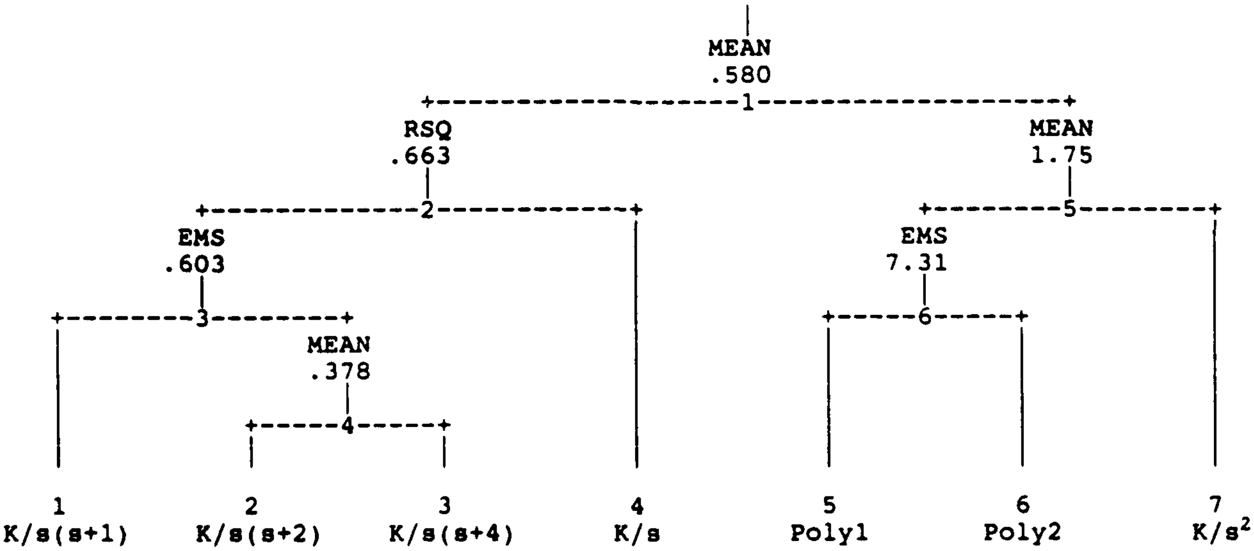
| Performance Measure | Relative Importance for Discriminating Between Systems |
|---|---|
| MEAN | 100 |
| EMS | 75 |
| MEAN_RUN | 74 |
| RMS | 72 |
| $P_3$ | 69 |
| COOPER-HARPER | 63 |
| RSQ | 61 |
| $P_2$ | 56 |
| $P_4$ | 49 |
| $P_1$ | 49 |

Note that the ordering of performance measures by CART is similar to that produced by STEPDISC (Section IV.C.2) in that MEAN, MEAN_RUN, and EMS are near the top of the list and the parameters associated with the control law, $P_1$-$P_4$, are nearer the bottom. The relative placement of RSQ and RMS, however, are different; STEPDISC places RMS at the bottom while CART ranks it higher in importance while the opposite is true for RSQ. Such discrepancies are expected because of the high correlation between some performance measures and the differences in the ranking criteria inherent in CART and STEPDISC.

On the basis of the combined information about variable importance from CART, STEPDISC, and CANDISC, CART was used to build and test a model for classification of the pilot runs based on three performance measures: MEAN, EMS, and RSQ. MEAN_RUN was dropped because of its high correlation with MEAN and because subsequent graphical interpretation of the output of CART is more easily discernible in three dimensions. RSQ was used in preference to RMS due to the very low importance found for RMS by CANDISC while CART shows comparable discrimination value for the two, albeit a preference for RMS. Several other groups of performance measures were considered but were found to produce similar results with less clarity of interpretation.

The resulting classification tree model is presented in Figure 27. This tree has seven terminal nodes; each terminal node is identified with one of the seven systems representing different handling qualities. These terminal nodes are defined through six splitting rules. Each splitting rule specifies a splitting variable and a splitting value. The splitting variables and values are referenced at the intermediate nodes in Figure 27. Pilot runs entering a splitting node are *split to the left branch* if their splitting variable value is less than or equal to the splitting value; otherwise, the run is split to the right branch.

**Figure 27.** Classification Tree Diagram
Tree with Seven Terminal Nodes



CART's options are set for determination of a cross-validated tree with ten-fold cross-validation. The cross-validation option causes CART to build and test classification tree models with various subsets of the data. Cross-validation is used to determine a misclassification rate for the final classification tree model. The selection of CART options also determines how many terminal classification nodes will be in the tree model.

Figure 28 presents the information available in CART output for each splitting rule. Splitting is done to reduce a cost parameter associated with misclassifying systems. The splitting variable at a node is selected to minimize the cost of misclassification for the resulting tree. Basically then, the split at each node attempts to separate systems. The output lists which variables are competitors for use as splitting variables at that node, together with the "improvement" that such a choice would bring to the tree. (See Appendix D.) In particular, a surrogate splitting variable strongly associated with the selected splitting variable is typically identified. This is a performance measure accomplishing a split of runs close to that provided by the actual splitting variable. The strength of association between the surrogate variable and the splitting variable is indicated.

**Figure 28.** Splitting Rule Information from CART Output

```
          *             Node 1 was split on variable MEAN
         * *            A case goes left if variable MEAN .le. 5.80e-01
        *   *           Improvement = 1.3e-01
       *   1 *
        *   *           Node     Cases     Class       Cost
         * *               1       168         1       0.86
          *                2       111         2       0.75
         * *               4        57         3       0.67
        *   *
      111    57         Class       Number Of Cases      Within Node Prob.
      *       *                     Top   Left  Right     Top  Left  Right
     *         *       K/s(s+1)     17     17     0      0.14  0.25  0.
    *           *      K/s(s+2)     18     18     0      0.14  0.25  0.
   * *         * *     K/s(s+4)     32     32     0      0.14  0.25  0.
  *   *       *   *    K/s          47     44     3      0.14  0.24  0.02
 *   2 *     *   4 *   Poly1        27      0    27      0.14  0.    0.33
  *   *       *   *    Poly2        11      0    11      0.14  0.    0.33
   * *         * *     K/s^2        16      0    16      0.14  0.    0.33
    *           *
                          Surrogate    Split          Assoc.  Improve.
                        1 RSQ       r  4.43e-01       0.62    6.8e-02

                          Competitor   Split                  Improve.
                        1 EMS          9.05e+00               8.7e-02
                        2 RSQ          4.98e-01               7.1e-02
```

The diagrams in Figures 29a through 29h show geometrically how the classification tree in Figure 27 defines regions of values of the three performance measures associated with the seven different systems. Note that all three performance measures are nonnegative. Also, RSQ is logically bounded above by one. MEAN and EMS actually take values larger than the limits on the axes in these figures; however, this only occurs for runs with the systems Poly2 and $K/s^2$. Thus, some runs with systems Poly2 and $K/s^2$ are out of the range of these plots and are thus not pictured in Figures 29g and 29h.

The first diagram, Figure 29a, shows how the seven CART classes partition the three-dimensional space formed by MEAN, EMS, and RSQ. The way that the performance measures separate systems into CART terminal nodes reiterates the system discrimination results reported earlier from the application of CANDISC. Specifically,

> MEAN separates the system into three groups:
> 1) K/s(s+1), K/s(s+2), K/s(s+4) and K/s
> 2) Poly1 and Poly2
> 3) $K/s^2$
>
> Within the first group, high RSQ values characterize K/s, and higher EMS values distinguish K/s(s+2) and K/s(s+4) from K/s(s+1).
>
> MEAN is used again to discriminate between K/s(s+2) and K/s(s+4).
>
> EMS distinguishes between the Poly1 and Poly2.

The diagrams in Figures 29b through 29h present the correctly classified runs from each system in the associated CART terminal node. These diagrams help to visualize the way that these performance measures characterize the systems and discriminate between the systems. The geometric details of the discrimination of systems through CART are quite straightforward and hence are more easily understood and interpreted than the discrimination via canonical variables in CANDISC. The ability to visualize CART's classification is an advantage as is CART's automatic allowance for heterogeneity with regards to the selection of variables to discriminate between different subsets of systems.

Tables 8 and 9 present how well this seven-node tree discriminates between the seven systems. Table 8 displays how the 168 pilot runs are classified by the CART tree. Each cell in Table 8 is associated with a "True System" and a "Predicted System". The cell contains the number of runs from the "True System" classified as the "Predicted System". There is also a fractional breakdown showing how the runs in each "True System" are distributed by the classification tree into the "Predicted Systems"; hence these fractions sum to one vertically in the table. Table 9, the cross-validation table, presents CART's estimated classification/misclassification probabilities for the seven-node tree; these probabilities are estimated by CART using cross-validation techniques which are described in Appendix D.

## Figure 29. Geometric View of CART Terminal Node Classes

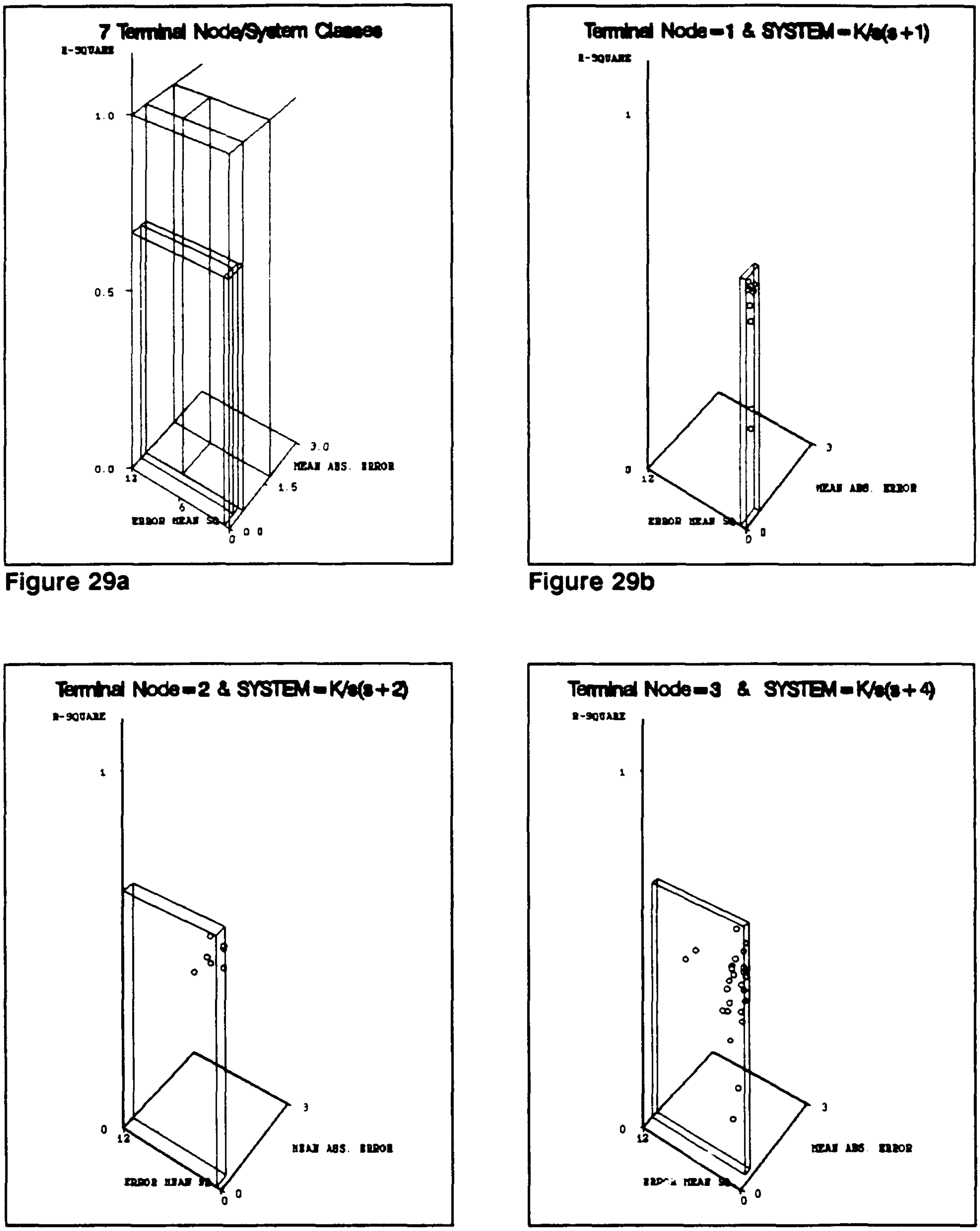


Figure 29a

Figure 29b

Figure 29c

Figure 29d

Figure 29. (Continued) Geometric View of CART Terminal Node Classes

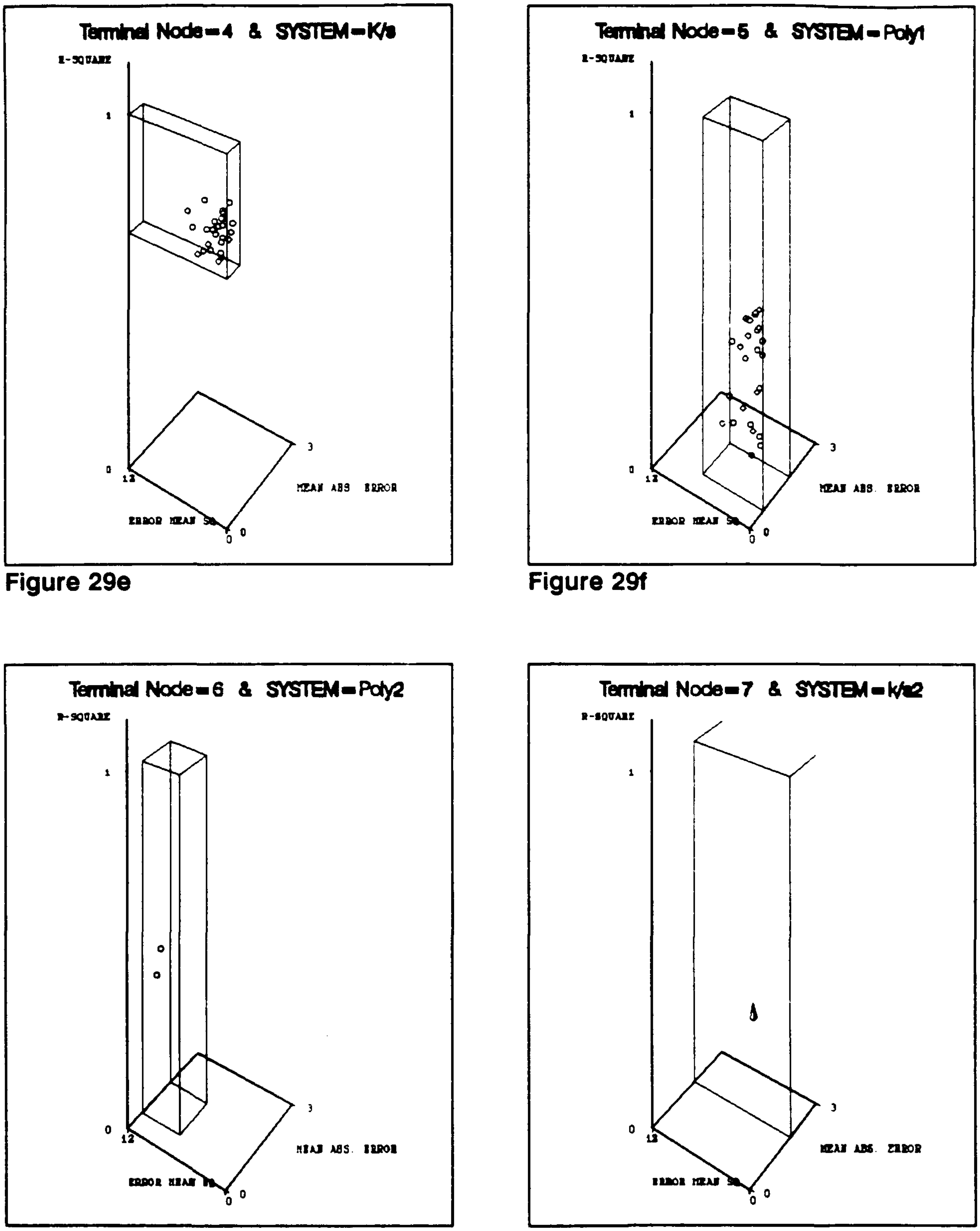


Figure 29e

Figure 29f

Figure 29g

Figure 29h

**Table 8.** Classification Matrix for Pilot Runs
Number of Runs (Fraction of Runs)

| Predicted System | True System<br>K/s(s+1) | K/s(s+2) | K/s(s+4) | K/s | Poly1 | Poly2 | $K/s^2$ |
|---|---|---|---|---|---|---|---|
| K/s(s+1) | 10 (.59) | 1 (.06) | 6 (.19) | 3 (.06) | 0 (.00) | 0 (.00) | 0 (.00) |
| K/s(s+2) | 1 (.06) | 7 (.38) | 0 (.00) | 3 (.06) | 0 (.00) | 0 (.00) | 0 (.00) |
| K/s(s+4) | 4 (.23) | 9 (.50) | 26 (.81) | 10 (.21) | 0 (.00) | 0 (.00) | 0 (.00) |
| K/s | 2 (.12) | 1 (.06) | 0 (.00) | 28 (.60) | 0 (.00) | 0 (.00) | 0 (.00) |
| Poly1 | 0 (.00) | 0 (.00) | 0 (.00) | 3 (.06) | 26 (.96) | 3 (.27) | 0 (.00) |
| Poly2 | 0 (.00) | 0 (.00) | 0 (.00) | 0 (.00) | 0 (.00) | 8 (.73) | 0 (.00) |
| $K/s^2$ | 0 (.00) | 0 (.00) | 0 (.00) | 0 (.00) | 1 (.04) | 0 (.00) | 16(1.0) |

**Table 9.** Cross-Validation Classification Matrix for Pilot Runs
Estimated Probability of Classification/Misclassification

| Predicted System | True System<br>K/s(s+1) | K/s(s+2) | K/s(s+4) | K/s | Poly1 | Poly2 | $K/s^2$ |
|---|---|---|---|---|---|---|---|
| K/s(s+1) | .35 | .06 | .25 | .06 | 0 | 0 | 0 |
| K/s(s+2) | .12 | .39 | .22 | .11 | 0 | .09 | 0 |
| K/s(s+4) | .41 | .44 | .53 | .17 | 0 | 0 | 0 |
| K/s | .12 | .11 | 0 | .60 | 0 | 0 | 0 |
| Poly1 | 0 | 0 | 0 | .06 | .96 | .18 | 0 |
| Poly2 | 0 | 0 | 0 | 0 | 0 | .73 | 0 |
| $K/s^2$ | 0 | 0 | 0 | 0 | .04 | 0 | 1 |

Both tables show that systems (and thus handling qualities) can be characterized and evaluated on the basis of these three performance measures. Specifically, the Poly1, Poly2, and $K/s^2$ systems are easily separated from the K/s and K/s(s+a), a=1,2,4, runs.

These measures have some problems telling Poly2 from Poly1 runs but still almost 75% of the Poly2 runs are correctly classified. At least half of the K/s and K/s(s+4) runs are correctly classified although there is some tendency for classifying K/s as K/s(s+4) and K/s(s+4) as K/s(s+1) and K/s(s+2). Generally, the K/s(s+a), a=1,2,4, runs are not that well separated showing the limitations of the performance measures for separating these systems.

Up to this point, the discrimination analysis of systems has ignored the potential effect of other sources of variation among runs besides systems. As noted in Section II.B, the individual pilot runs differ not only in the system used but also in four additional design parameters. These parameters are gain, vertical display scale of the screen, the bandwidth controlling the pitch of the tracking target, and the presence of the secondary loading task. These design parameters are described in detail in Section II.A. Gain is a multiplier in the system transfer function and thus models secondary variations in handling qualities. Each system was run at multiple levels of gain. The screen vertical display scale determines the dimensions of the tail-chase scene presented to the pilot. Two vertical display scales were used: "normal" and wide". The bandwidth parameter determines how many of the input signal sinusoids would be high amplitude sinusoids and thus changes the tracking task performed. Two bandwidths were used: "low" and "high". Finally, each system was run with the secondary task off and with the secondary task on. This, like bandwidth, changes the tracking task performed.

To investigate whether these design parameters explain some of the mixing of systems across terminal nodes leading to misclassification, *Tables 10 through 16 break down the* pilot runs for each system according to the CART terminal node assigned to the run and also according to the settings of the four design parameters. In these tables, both the number of the terminal node (as numbered in Figure 27) and the system identified with the terminal node are listed. Unfortunately, the small number of runs in these tables at any specific configuration of the four design parameters preclude definitive conclusions being reached about the association between these four design parameters and misclassification. However, the pattern of misclassified runs suggests that higher gains and the presence of the secondary task may lead to misclassification of one system as another in some cases, as detailed below.

## **Table 10.** Breakdown of Pilot Runs by CART Terminal Node and Design Parameters for SYSTEM = K/s(s + 1)

| | Total Classi-fications | System Gain | | | | Bandwidth | | Screen | | Secondary Task | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Terminal Node | | 17.60 | 35.20 | 58.60 | 83.80 | Low | High | Normal | Wide | Off | On |
| 1: K/s(s+1) | 10 | | 8 | 1 | 1 | 7 | 3 | 8 | 2 | 1 | 9 |
| 2: K/s(s+2) | 1 | | 1 | | | 1 | | 1 | | | 1 |
| 3: K/s(s+4) | 4 | 1 | 3 | | | 2 | 2 | 4 | | | 4 |
| 4: K/s | 2 | | 2 | | | 2 | | 2 | | 2 | |

Table 10 does not indicate any pattern of misclassification of system K/s(s + 1) related to the design parameters.

## **Table 11.** Breakdown of Pilot Runs by CART Terminal Node and Design Parameters for SYSTEM = K/s(s + 2)

| | Total Classi-fications | System Gain | | | | | Bandwidth | | Screen | | Secondary Task | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Terminal Node | | 8.38 | 17.60 | 21.50 | 35.20 | 58.60 | Low | High | Normal | Wide | Off | On |
| 1: K/s(s+1) | 1 | | | | 1 | | | 1 | | 1 | | 1 |
| 2: K/s(s+2) | 7 | | 1 | 2 | 4 | | 7 | | 6 | 1 | 3 | 4 |
| 3: K/s(s+4) | 9 | 1 | 1 | 1 | 5 | 1 | 8 | 1 | 9 | | 1 | 8 |
| 4: K/s | 1 | | | | 1 | | 1 | | 1 | | 1 | |

Eight of the nine system K/s(s + 2) runs misclassified as system K/s(s + 4) in Table 11 have the secondary task on. Also, the one run at the highest gain value (58.60) is associated with this misclassification.

## **Table 12.** Breakdown of Pilot Runs by CART Terminal Node and Design Parameters for SYSTEM = K/s(s + 4)

| | Total Classi-fications | System Gain | | | | | Bandwidth | | Screen | | Secondary Task | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Terminal Node | | 17.60 | 35.20 | 58.60 | 83.80 | 117 | Low | High | Normal | Wide | Off | On |
| 1: K/s(s+1) | 6 | | | 3 | 1 | 2 | 5 | 1 | 3 | 3 | | 6 |
| 3: K/s(s+4) | 26 | 2 | 12 | 12 | | | 23 | 3 | 25 | 1 | 7 | 19 |

Table 12 indicates that secondary task on and high gain may lead to the misclassification of system K/s(s+4) as K/s(s+1).

**Table 13.** Breakdown of Pilot Runs
by CART Terminal Node and Design Parameters
for SYSTEM=K/s

| | Total Classi-fications | System Gain | | | | | Bandwidth | | Screen | | Secondary Task | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Terminal Node | | 0.29 | 0.59 | 1.17 | 5.86 | 8.38 | Low | High | Normal | Wide | Off | On |
| 1: K/s(s+1) | 3 | | | | 3 | | 3 | | 3 | | 2 | 1 |
| 2: K/s(s+2) | 3 | | 3 | | | | 3 | | 3 | | 1 | 2 |
| 3: K/s(s+4) | 10 | 1 | 8 | | | 1 | 10 | | 9 | 1 | 1 | 9 |
| 4: K/s | 28 | 4 | 23 | 1 | | | 26 | 2 | 28 | | 16 | 12 |
| 5: Poly1 | 3 | 1 | 2 | | | | 2 | 1 | 2 | 1 | | 3 |

Table 13 indicates that secondary task on and high gain values can lead to the misclassification of system K/s runs as K/s(s+a), a=1,2,4, systems.

**Table 14.** Breakdown of Pilot Runs
by CART Terminal Node and Design Parameters
for SYSTEM=Poly1

| | Total Classi-fications | System Gain | | | | | Bandwidth | | Screen | | Secondary Task | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Terminal Node | | 17.60 | 21.50 | 35.20 | 58.60 | 83.80 | Low | High | Normal | Wide | Off | On |
| 5: Poly1 | 26 | 7 | 7 | 8 | 1 | 3 | 21 | 5 | 20 | 6 | 6 | 20 |
| 7: K/$s^2$ | 1 | | | 1 | | | 1 | | 1 | | | 1 |

Since only one run is misclassified, no pattern of misclassification can be determined from Table 14.

**Table 15.** Breakdown of Pilot Runs
by CART Terminal Node and Design Parameters
for SYSTEM = Poly2

| | Total Classi- | System Gain | | | Bandwidth | | Screen | | Secondary Task | |
|---|---|---|---|---|---|---|---|---|---|---|
| Terminal Node | fications | 35.20 | 83.80 | 215 | Low | High | Normal | Wide | Off | On |
| 5: Poly1 | 3 | 1 | 1 | 1 | 3 | | 2 | 1 | | 3 |
| 6: Poly2 | 8 | 8 | | | 7 | 1 | 8 | | 1 | 7 |

Table 15 indicates that secondary task on and high gain values can lead to misclassification of system Poly2 as system Poly1.

**Table 16.** Breakdown of Pilot Runs
by CART Terminal Node and Design Parameters
for SYSTEM = $K/s^2$

| | Total Classi- | System Gain | | | Bandwidth | | Screen | | Secondary Task | |
|---|---|---|---|---|---|---|---|---|---|---|
| Terminal Node | fications | 0.29 | 0.59 | 1.17 | Low | High | Normal | Wide | Off | On |
| 7: $K/s^2$ | 16 | 4 | 11 | 1 | 12 | 4 | 13 | 3 | 1 | 15 |

All pilot runs using $K/s^2$ are correctly classified as shown in Table 16.

## D. Evaluation of Handling Qualities through the Analysis of Performance Measures

Each flight simulation system used in this experiment represents a category of handling qualities. The preceding analysis in Section IV.C explores the ability of the performance measures to discriminate these systems. In other words, that analysis concentrates on demonstrating that there is a relationship between objective performance measures and systems which allows systems to be classified according to the values taken on by performance measures in associated pilot runs. This section tries to demonstrate that through this relationship, the performance measures provide meaningful insight for the evaluation of handling qualities. Accordingly, it focuses on the type of information that the performance measures provide.

First, consider the statistical error measures EMS and RSQ and mean absolute tracking error measure MEAN. Recall that these are the measures singled out by CANDISC and used subsequently in the CART analysis. The physical/qualitative meanings of the three performance measures are:

EMS measures the variation of run data around the estimated control law; that is, it tells how closely the pilot was able to maintain his ideal control law.

RSQ measures how much control information is in the estimated control law; that is it tells how well the estimated control law describes actual pilot control input.

MEAN measures the accuracy of the estimated control law for performing the tracking task.

These performance measures are objective in that they are analytically derived from objectively measured pilot response. They express the consistency with which the pilot "flies" as well as how accurately the tracking task is performed.

Note that EMS=0 is equivalent to RSQ=1. That is, the estimated control law exactly fits the measured pilot response when and only when the residual errors are all zero. If the estimated control law performs the pitch tracking task with zero error in the closed loop simulation run, then MEAN will be zero. From an analysis viewpoint then, the desirable values for these performance measures are EMS=0, RSQ=1, and MEAN=0. In that case, the pilot would have perfectly defined and perfectly "flown" a control law that executed the tracking task without error.

In this context, each of the systems can be evaluated and compared relative to how pilot runs using that system deviate from this desired state. For example, referring to Figure 29a:

The runs using system K/s have the highest RSQ values while maintaining low MEAN values. Thus, the system K/s runs show the most clearly defined control laws and perform the tracking task well.

The system K/s(s+a), a=1,2,4, runs demonstrate lower RSQ values than the system K/s runs and thus have less well-defined control laws than for K/s runs; but, they perform the tracking task as well as is demonstrated by having similar values of MEAN. Within the K/s(s+a) systems for varying "a", the system K/s(s+2) runs have lower tracking error (smaller MEAN values) than K/s(s+4), and control inputs in system K/s(s+1) runs show the least variation around the estimated control laws since they have smaller EMS values.

Both Poly system runs show higher tracking error (larger MEAN values) than either K/s or K/s(s+a), a=1,2,4, runs. Also, the control inputs in system Poly1 runs show smaller variation around the estimated control laws (smaller EMS values) than Poly2 runs.

The K/$s^2$ runs demonstrate the worst handling qualities of all seven systems considered as shown by having the worst values for EMS, RSQ, and MEAN.

As discussed in Section III.B, an important intermediate analysis product in the calculation of the error measures and tracking errors is the estimation of the pilot's control law. The statistical process of estimating the control law yields the error measures EMS and RSQ, and the estimated control law used as an autopilot yields the task accuracy measure MEAN. Below, the estimated control law itself is investigated as to its role in the evaluation of handling qualities. Specifically, the four secondary performance measures $P_1$-$P_4$ introduced in Section III.D which characterize the geometric shape of the control law will be considered.

The estimated control laws for the correctly classified experimental runs at each node of the CART classification tree defined in Figure 27 are plotted in Figure 30. Each sub-figure presents overlaid plots of the estimated control laws for the different runs made using the indicated system. Here, as in some analysis to follow, only the correctly classified runs according to the CART classification tree are used to eliminate some of the aberrant runs which are not characteristic of other runs using the same system. As noted before, some runs differ due to variations in the design parameters (gain, screen scale, bandwidth, and secondary task) while other runs differ due to pilot differences. Removing these aberrant runs through the use of the CART analysis helps to clarify the utility of performance measures for evaluating handling qualities by reducing the variation in the performance measures not due to system differences.

Several differences in the overall shapes of the controls laws can be seen in Figure 30. For example, the inputted control is zero for zero tracking error except for the $K/s^2$ runs where the inputted control is positive. $P_1$ is the performance measure characterizing the inputted control for zero tracking error so this difference in $K/s^2$ runs can also be seen in Figure 23. In Figure 23, the median $P_1$ values are slightly negative for all runs except those using $K/s^2$ for which the median is much larger and positive. Ideally, the inputted control for zero tracking error should be zero. This tendency for being positive for $K/s^2$ runs may be related to some anticipation by the pilot or to some undesirable bias due to overcompensation when using this system.

Another difference in the control laws is their symmetry which appears visually in Figure 30 as the crossing point of the overlaid control laws. Symmetry is characterized by the performance measure $P_2$. Recall that $P_2$ is the tracking error at the center of the estimated control law curve and thus measures any difference in how the pilot performs when inputting positive and negative controls. In particular, it measures asymmetry with respect to zero tracking error of the range over which the control law is linear and any asymmetry in the tendency to saturate the control law. As further seen in Figure 24, both $K/s(s+1)$ and $K/s^2$ system runs have a tendency to negative values of $P_2$ while the other systems have more positive values. Thus, for $K/s(s+1)$ and $K/s^2$ runs, the pilot tends to saturate his inputted control more for positive inputs than negative and his control law is more linear for negative inputted controls than for positive. The opposite tendency holds for the runs from the other systems.

**Figure 30.** Estimated Control Laws for Correctly Classified Runs

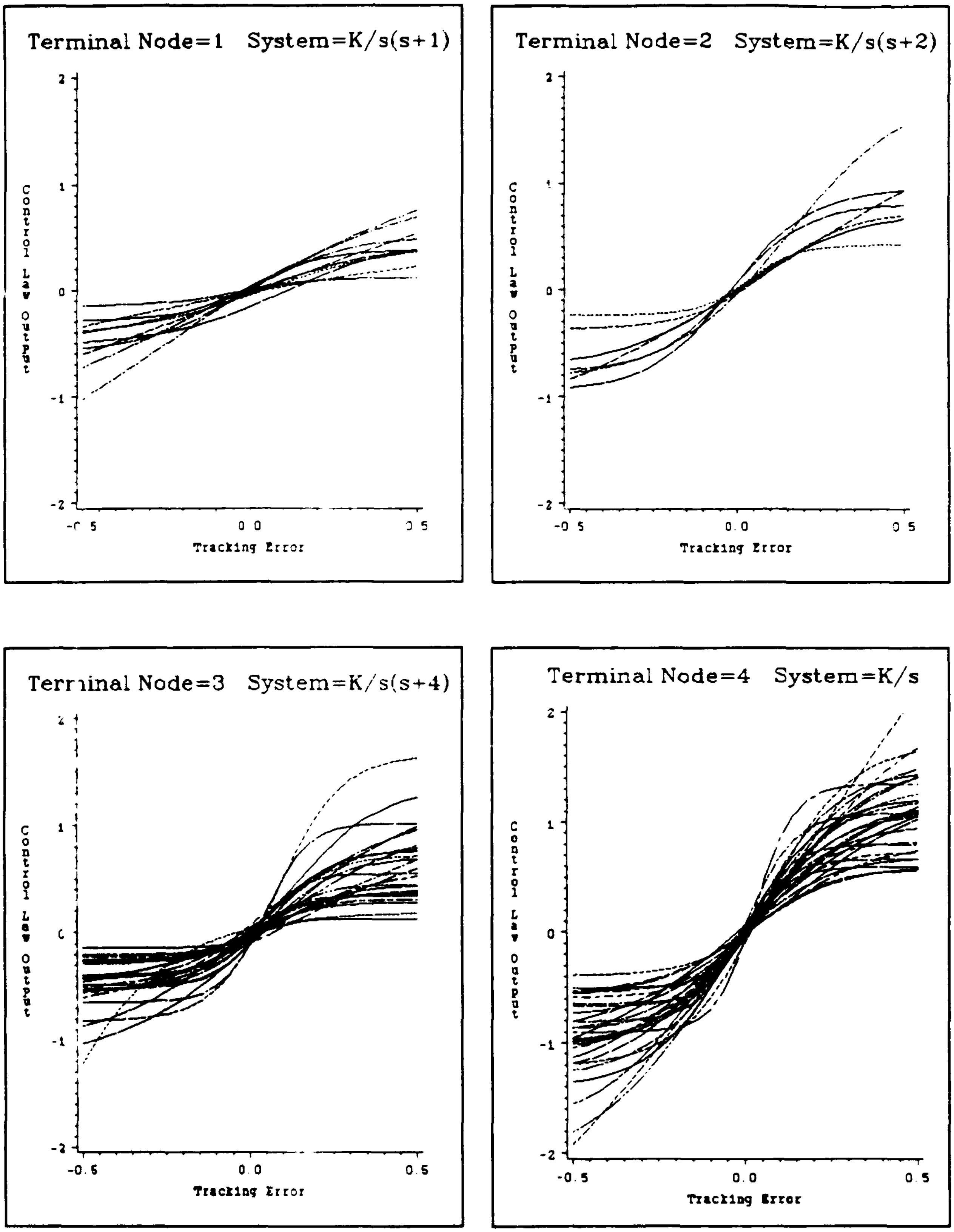
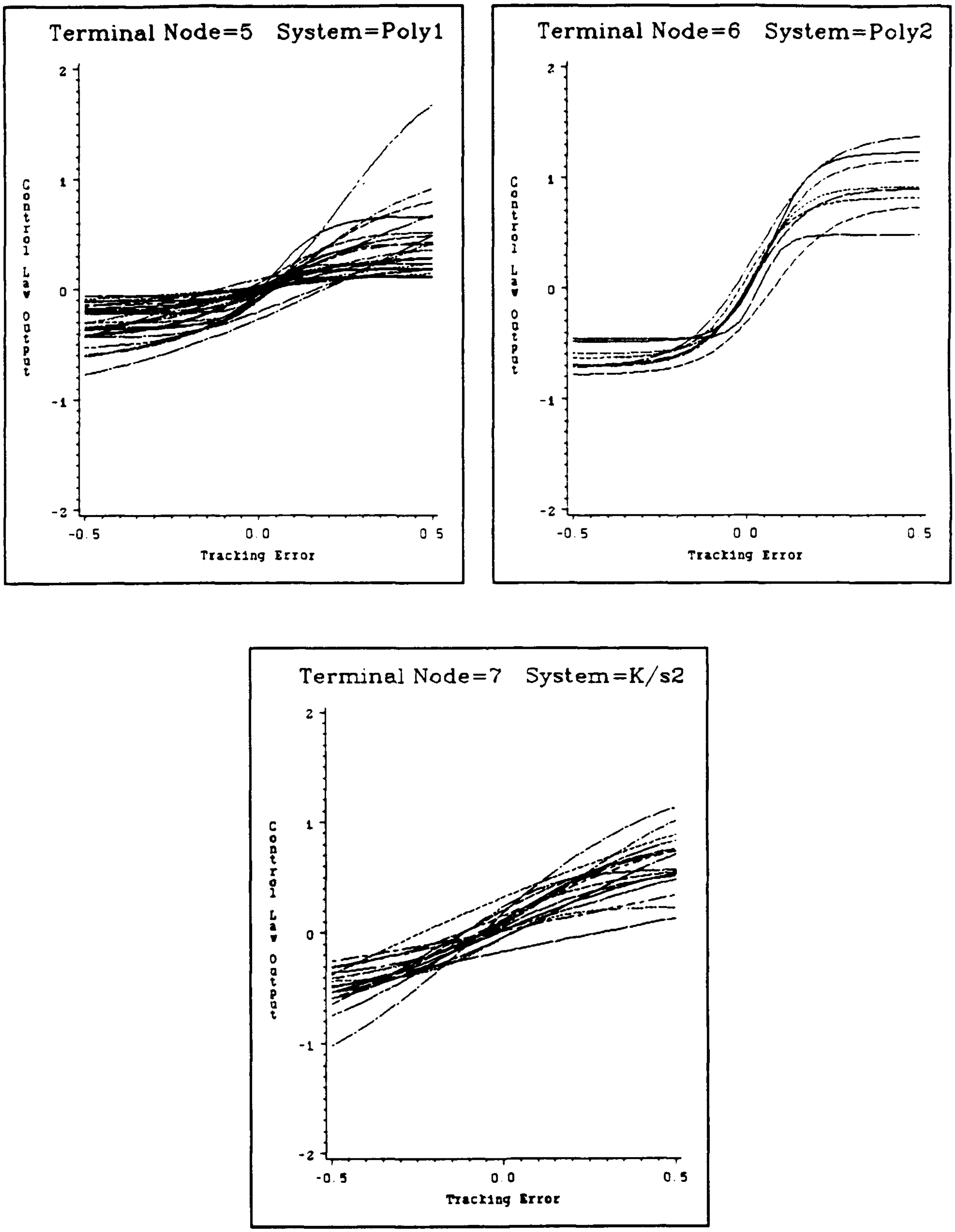


**Figure 30.** (Continued) Estimated Control Laws for Correctly Classified Runs


Terminal Node=5 System=Poly1

Terminal Node=6 System=Poly2

Terminal Node=7 System=K/s2

The most pronounced difference in the control laws shown in Figure 30 is the amount of curvature. The control laws for systems K/s(s+1), Poly1, and $K/s^2$, are "flatter" than the control laws for the other systems; in addition, the control laws for systems K/s(s+1) and $K/s^2$ are more linear. This difference in curvature can more clearly be seen in Figure 31 showing overlays of the median control laws for the different systems.

The performance measure characterizing the linearity of the control law is $P_4$, the tracking error interval over which the estimated control law curve is linear. As noted in Figure 26, $K/s^2$ and K/s(s+1) have the largest value of this parameter; the median values for the remaining systems are very similar. Of course, the more linear a control law is, the less the pilot saturates his inputted control. Thus, for these two systems, the pilots did not have as much tendency to ease off of control inputs for larger tracking errors as for the remaining systems.

The performance measure characterizing control rate, i.e., the slope of the "middle" part of the control law, is $P_3$. The hierarchical arrangement of system by control rates observed in Figures 30 and 31 can be seen again using the box plots in Figure 32 summarizing the parameter $P_3$ for the correctly classified pilot runs. The presentation in this figure is similar to the presentation in Figure 25 which summarizes all 168 experimental runs. However, by using only the correctly-classified runs, Figure 32 provides a better discrimination of systems than Figure 25, thus demonstrating the utility of removing aberrant runs as discussed earlier.

**Figure 32.** Control Law Parameter $P_3$
Box Plot Summary by System and CART Terminal Node
Correctly Classified Pilot Runs Only

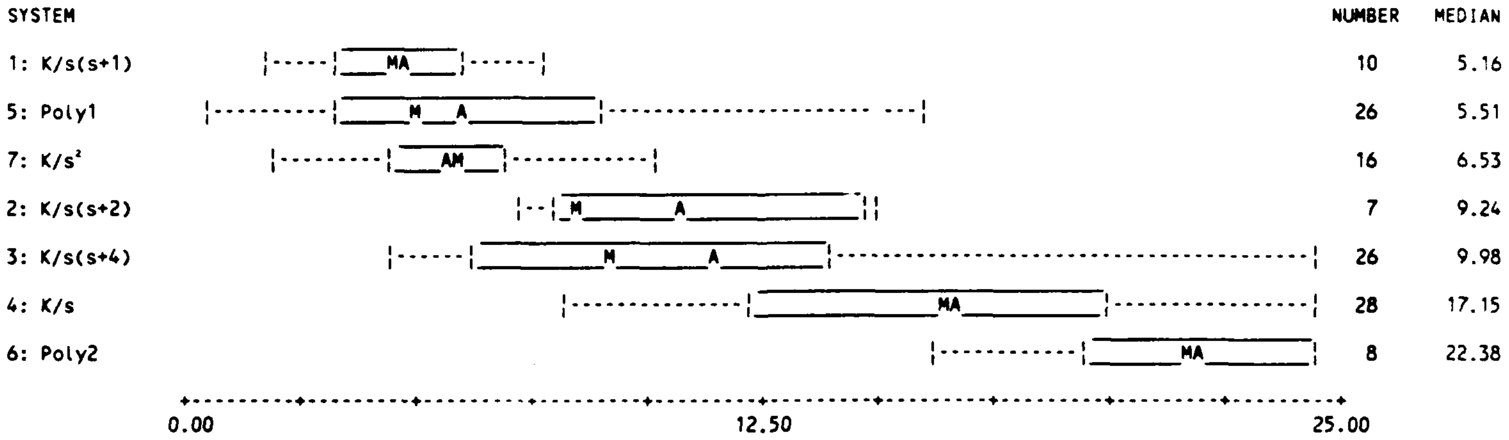
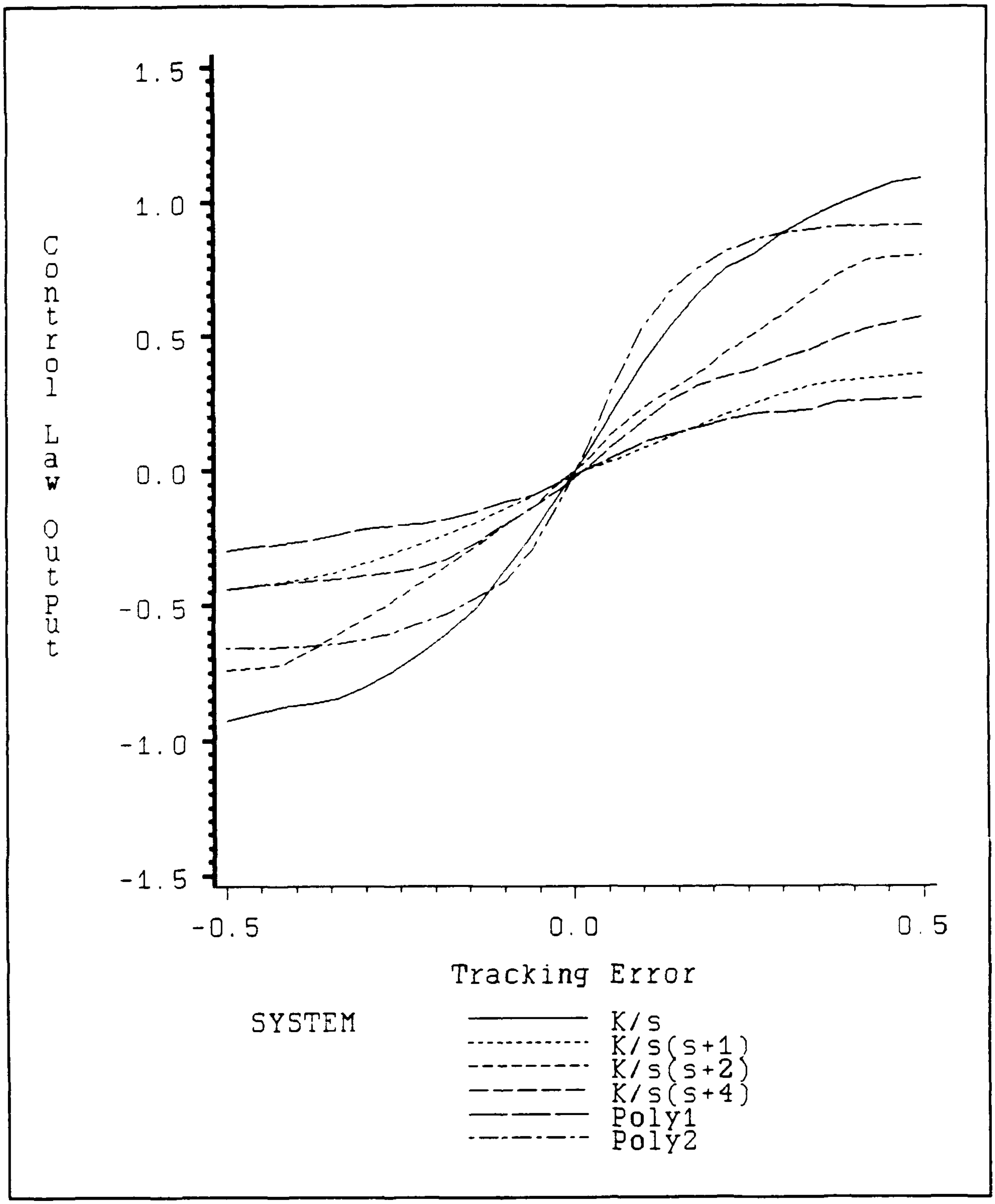


**Figure 31.** Median Control Laws for Each System (Excluding K/s$^2$)



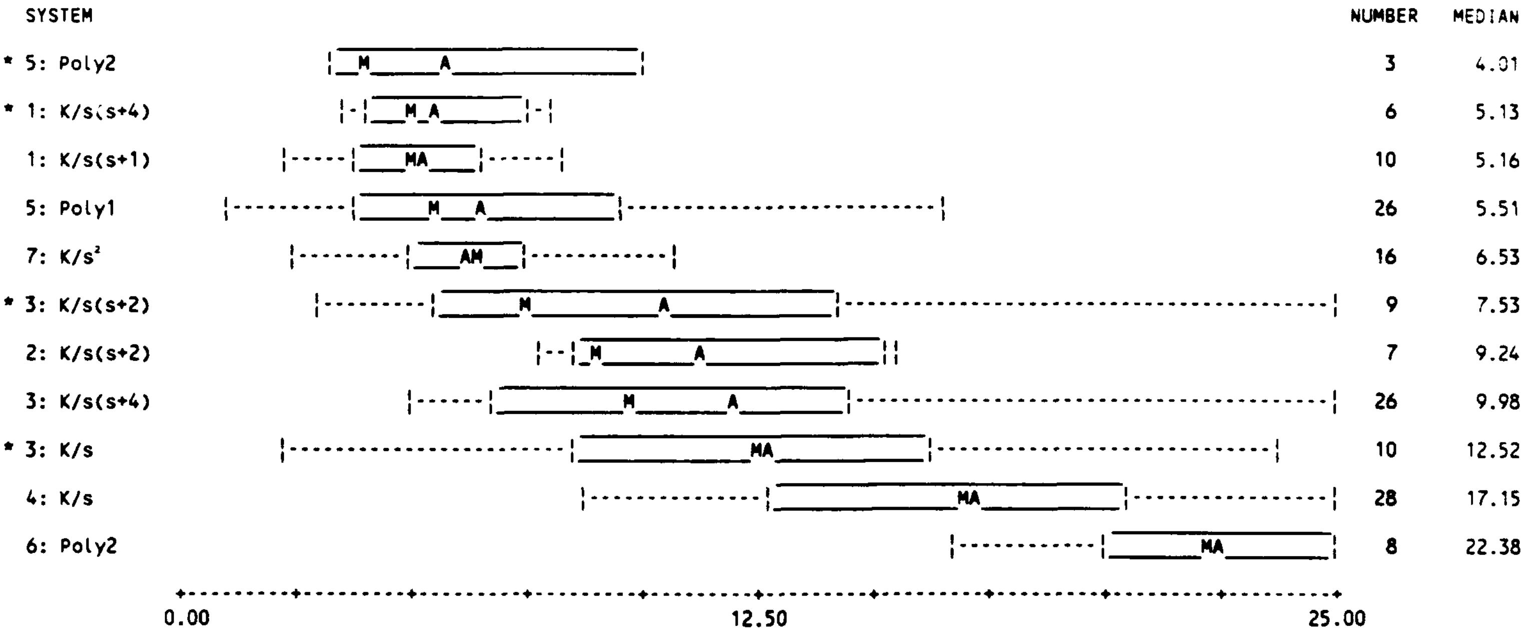
The control rate is related to handling qualities in that the control rate is largely determined by the responsiveness of a system with a more responsive system showing a smaller control rate. Figure 32 indicates low control rates for systems K/s(s+1), Poly1 and K/s$^2$, moderate control rates for systems K/s(s+2) and K/s(s+4), a higher rate for system K/s, and a still higher rate for system Poly2. Thus, K/s(s+1), Poly1, and K/s$^2$ are the most responsive systems followed by K/s(s+2) and K/s(s+4) with K/s and Poly2 being the least responsive systems investigated.

Figure 33 augments the summary in Figure 32 by adding box plots for four groups of runs misclassified by CART. These four groups are:

9 K/s(s+2) runs misclassified as K/s(s+4)<br>
6 K/s(s+4) runs misclassified as K/s(s+1)<br>
10 K/s runs misclassified as K/s(s+4)<br>
3 Poly2 runs misclassified as Poly1

These four groups represent the largest groups of misclassified runs for each system. Groups from K/s(s+1), Poly1, and K/s$^2$ are not included due to the small number of misclassifications for these systems.

**Figure 33.** Control Law Parameter $P_3$
Box Plot Summary by System and CART Terminal Node
Misclassified Groups Indicated by '*'



All misclassifications result in lower values of $P_3$, that is, in lower control rates. As noted in Section IV.C.3, all of the misclassifications included in Figure 33 are associated with the following design parameter settings: secondary task on and high gain values. It is reasonable that a higher gain and/or a demanding secondary task would reduce the pilot's control rate.

The auxiliary information from parameter $P_3$ related to the misclassifications can be integrated with the information from the CART classification analysis to further qualify the handling qualities embodied in the systems. Four of the systems have *primary* runs associated with that system's CART terminal node and *misclassified* runs associated with another system's CART terminal node. It was just noted that $P_3$ is smaller for the *misclassified* runs than for the *primary* runs.

> The K/s *misclassified* runs are in CART terminal node 3 which has lower RSQ values than the *primary* K/s runs in terminal node 4. The *misclassified* K/s runs then are differentiated by less pilot consistency as well as by lower pilot control rates.
>
> The K/s(s+2) *misclassified* runs are in CART terminal node 3 which has higher MEAN values then the *primary* K/s(s+2) runs in terminal node 2. The *misclassified* K/s(s+2) runs then are differentiated by less accuracy in task performance as well as by lower pilot control rates.
>
> The K/s(s+4) *misclassified* runs are in CART terminal node 1 which has lower EMS values than the *primary* K/s(s+4) runs in terminal node 3. The *misclassified* K/s(s+4) runs then are differentiated by a closer pilot adherence to the estimated control law as well as by lower pilot control rates.
>
> The Poly2 *misclassified* runs are in CART terminal node 5 which has lower EMS values than the *primary* Poly2 runs in terminal node 6. The *misclassified* Poly2 runs then are differentiated by a closer pilot adherence to the estimated control law as well as by lower pilot control rates.

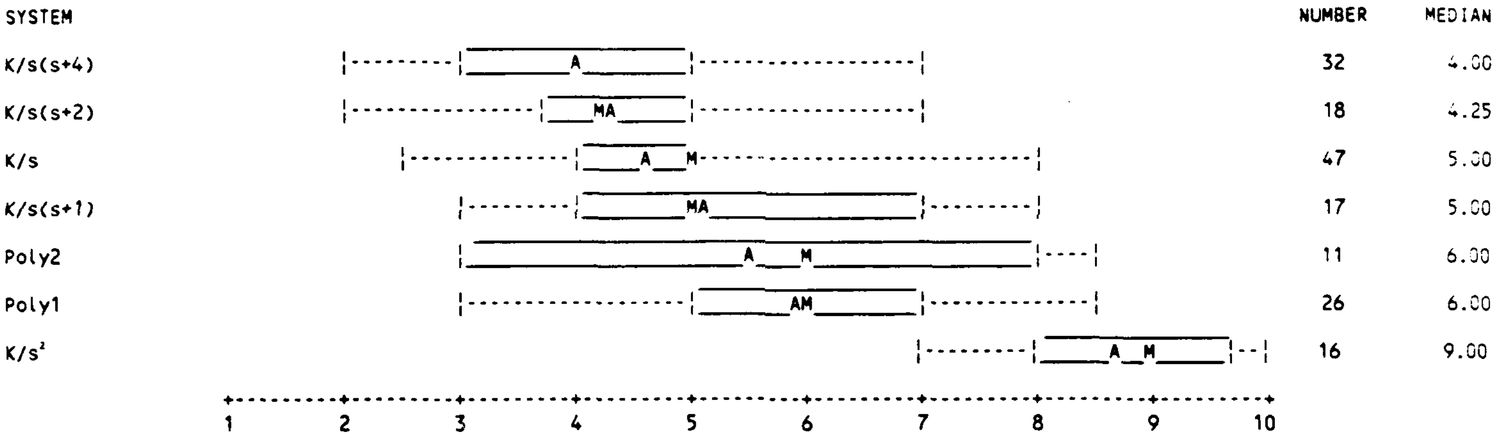
These observations seem to imply that the K/s and K/s(s+2) *primary* runs exhibit preferable handling qualities to the *misclassified* runs. But with the K/s(s+4) and Poly2 systems, the *misclassified* runs exhibit preferable handling qualities. These evaluations are based on the interpretations of the CART terminal nodes in terms of levels of MEAN, EMS, and RSQ. Since the main difference in handling qualities reflected in the misclassified runs is higher gain, there is some indication that a higher gain may be preferable for the systems K/s(s+4) and Poly2. Since the misclassified runs are also associated with the secondary task on, this preference for higher gain may also be only associated with the more complicated tracking task.

# SECTION V
# RELATIONSHIP BETWEEN
# COOPER-HARPER RATING AND PERFORMANCE MEASURES

## A. Subjective Pilot Opinion - Cooper-Harper Rating

The analyses discussed so far concerning evaluation of handling qualities do not address pilot opinion of handling qualities as embodied in the Cooper-Harper rating described in Appendix C. All of the analyses have been based solely on performance measures derived from directly measured pilot response. Figure 34 summarizes the Cooper-Harper ratings for the seven experimental systems.

**Figure 34.** Cooper-Harper Rating
Box Plot Summary by Systems



The box plots in Figure 34 indicate an ordering of the systems by the Cooper-Harper rating although there is significant overlap in the ranges of Cooper-Harper ratings for most systems. The lowest Cooper-Harper ratings are for systems K/s(s+4) and K/s(s+2) which are separated from those for the "worst" systems Poly1 and K/s$^2$. The Cooper-Harper ratings for the remaining systems K/s, K/s(s+1), and Poly2 somewhat overlap the best and worst systems. Thus, unlike using the performance measures, there is no clear-cut division of systems by Cooper-Harper ratings alone. However, an advantage of Cooper-Harper is that it does provide a preference ordering of the systems.

## B. Correlation Between Cooper-Harper Rating and Performance Measures

Table 17 lists the correlations between Cooper-Harper rating and the performance measures.

**Table 17.** Correlation Between Cooper-Harper Rating and Performance Measures

| Performance Measure | Correlation |
|---|---|
| MEAN_RUN | .694 |
| MEAN | .555 |
| RSQ | -.392 |
| $P_3$ | -.337 |
| $P_4$ | .318 |
| EMS | .268 |
| $P_1$ | .177 |
| $P_2$ | -.066 |
| RMS | .058 |

As seen by the correlations, there is a fairly strong relationship between MEAN_RUN/MEAN and the Cooper-Harper rating. Also, note that the correlation with $P_3$ and $P_4$ exceeds that with EMS and RMS suggesting that the Cooper-Harper rating is related to the shape of the control law more so than to some of the error measures, contrary to the relationship between system and performance measures.

To further investigate the relationship between Cooper-Harper and performance measures, the discrimination algorithms STEPDISC and CANDISC were used with Cooper-Harper rating as the grouping variable. To eliminate many small groups, the Cooper-Harper rating was rounded to the nearest integer to produce nine groups ranging from 2 to 10.

STEPDISC using the STEPWISE option with significance level 0.15 for a variable to enter the model and to stay produced the following ranking:

$$\text{MEAN\_RUN} \rightarrow P_4 \rightarrow P_2 \rightarrow \text{MEAN} \rightarrow \text{RSQ} \rightarrow P_1 \rightarrow \text{EMS}$$

$P_3$ and RMS are not selected under these enter and stay levels. $P_3$ loses much of its discriminating power because of a strong correlation with MEAN_RUN. Table 18 shows the squared partial correlations as well as the squared correlations. Recall that the squared correlation is an indication of the performance measure's isolated ability to discriminate between Cooper-Harper ratings, whereas the squared partial correlation is a measure of the additional discrimination between Cooper-Harper ratings when the performance measure is added.

**Table 18.** Discrimination Power of Performance Measures Measured by Squared Correlation and Squared Partial Correlation with Cooper-Harper Rating (Rounded)

| Performance Measure | Sq. Partial Correlation | Squared Correlation |
|---|---|---|
| MEAN_RUN | .627 | .627 |
| $P_4$ | .210 | .310 |
| $P_2$ | .184 | .175 |
| MEAN | .176 | .493 |
| RSQ | .129 | .217 |
| $P_1$ | .105 | .158 |
| EMS | .102 | .223 |
| RMS | .060 | .080 |
| $P_3$ | .049 | .140 |

Table 18 suggests that MEAN_RUN, MEAN, $P_2$, and $P_4$ are the "main" performance measures associated with the Cooper-Harper rating. Unlike with systems, the Cooper-Harper rating is more related to perceived pilot error as embodied in MEAN_RUN as opposed to MEAN and is more related to the parameters governing the control law such as $P_2$ and $P_4$ than with error measures such as EMS, RSQ, and RMS.

CANDISC was also applied to the (rounded) Cooper-Harper rating. As expected from the analysis above, the first canonical variable is mostly related to MEAN_RUN (correlation=.918) with the next most related variables being MEAN (correlation=.785) and $P_4$ (correlation=.548). The total squared correlation between the Cooper-Harper rating and the canonical variable is .857 indicating the Cooper-Harper rating can be predicted with a high degree of accuracy using the associated values of the performance measures. The total squared correlation is comparable to that between system and performance measures as given in Table 5.

The pairwise tests of differences between Cooper-Harper ratings show that the lower values of Cooper-Harper are not well separated. Specifically, the values of the Cooper-Harper ratings fall roughly into three sets:

Set 1: 2, 3, 4<br>
Set 2: 5, 6, 7<br>
Set 3: 8, 9, 10

with the first two sets overlapping more so than the latter two. The linear ranking based on the first canonical variable also points out this grouping. Table 19 gives the median values of the canonical variable with each group.

**Table 19.** Medians of the Canonical Variable per Cooper-Harper Group

| Cooper-Harper Rating (Rounded) | Median Canonical Variable |
|---|---|
| 2 | -1.055 |
| 3 | -0.923 |
| 4 | -0.878 |
| 5 | -0.679 |
| 6 | -0.478 |
| 7 | 0.054 |
| 8 | 1.413 |
| 9 | 3.942 |
| 10 | 6.529 |

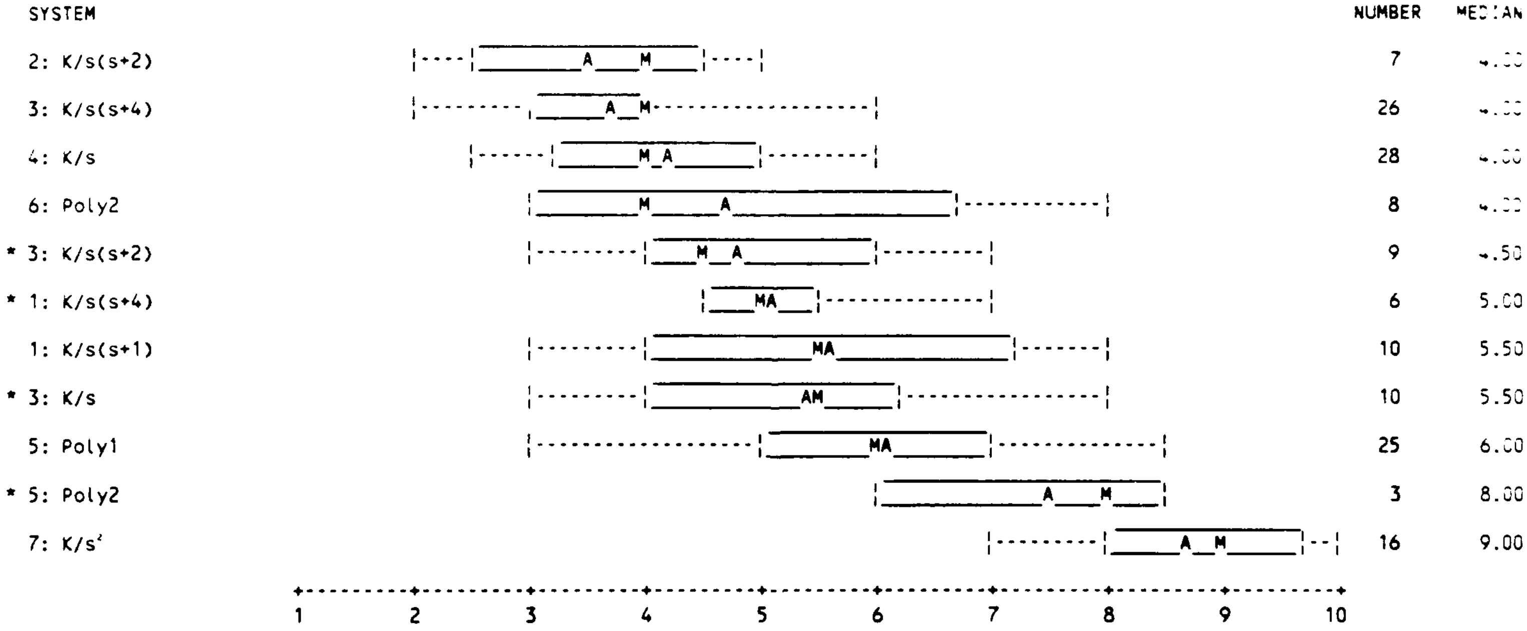
Note that the medians of the canonical variables are increasing with Cooper-Harper rating so that not only can performance measures be used to predict the Cooper-Harper rating but the relative value of the canonical variable is also a measure of preference of systems.

## C. Evaluation of Handling Qualities

Cooper-Harper ratings are related to pilot response as shown in the previous section which studied the correlation between Cooper-Harper ratings and the performance measures summarizing pilot response. The following discussion exemplifies how this relationship can enhance the Cooper-Harper subjective evaluation of handling qualities through integration with the performance measures of pilot response.

Figure 35 is a refinement of Figure 34; it summarizes the Cooper-Harper ratings for the eleven groups of runs treated in Figure 33. These eleven groupings consist of the seven *primary* system groups defined by the seven CART analysis terminal nodes in Figure 27 and four more *misclassified* system groups.

**Figure 35.** Cooper-Harper Rating
Box Plot Summary by System and CART Terminal Node
Misclassified Groups Indicated by '*'



Compared to the pattern in Figure 34, the *primary* system groups from the best systems, K/s(s+2) and K/s(s+4), are more cleanly separated from those of the worst systems, Poly1 and K/s$^2$. *Primary* system groups for the remaining systems, K/s, Poly2, and K/s(s+1), still lay in the midrange Cooper-Harper ratings somewhat overlapping the best and worst systems. This is another example of how relationships between variables is clarified via removal of aberrant runs using the CART classification tree in Figure 27.

Figure 33 shows how pilot control rates differ across the seven *primary* system groups. Comparing this figure to Figure 35, the two best Cooper-Harper rated systems, K/s(s+2) and K/s(s+4), have control rates in the midrange (9 to 10). The next best Cooper-Harper rated systems, K/s and Poly2, have significantly higher pilot control rates, 17 and 22, respectively. The worst Cooper-Harper rated systems, K/s(s+1), Poly1, and $K/s^2$, have low control rates, in the 5 to 6 range. Thus, there appears to be a middle range of control rates associated with preferred systems (handling qualities). Systems resulting in smaller or higher control rates are not preferred as measured by the Cooper-Harper rating with larger control rates being less detrimental than those with smaller rates. Linking control rates to responsiveness of the system, the resulting conclusion about handling qualities is that a highly responsive system is less preferable than a sluggish one with a moderate amount of responsiveness being preferable over either extreme.

This link between control rate and Cooper-Harper rating is further exemplified through comparing the *misclassified* with the *primary* system groups. From Figure 35, *primary* system groups for K/s(s+2), K/s(s+4), K/s, and Poly2 have median Cooper-Harper rating of 4. The *misclassified* system groups for each of these systems have significantly higher Cooper-Harper ratings. Earlier analyses showed that these *misclassified* system groups are all associated with lower control rates than the corresponding *primary* system groups. So, once again lower control rates are associated with deterioration of pilot opinion of the system.

# SECTION VI
# CONCLUSIONS

This research project is a demonstration of principles develped in earlier methodological work of Baldwin and Gantz (1983, 1984). A group of systems representing a variety of handling qualities are flown by test pilots. Objective dynamic information and measurements of pilot response are collected for the experimental runs and pilot opinion is recorded. By modeling pilot control via statistical estimation and joining this with an evaluation of the model's tracking performance, the systems, and by implication the handling qualities, are characterized by objective performance measures of pilot response. Additionally, the objective performance measures of pilot response are related to subjective measures of pilot opinion (such as Cooper-Harper ratings), and in fact are shown to be predictive of pilot opinion.

Procedures followed in this research are generalizable to other experimental flight simulations and flight test environments. Elements of the statistical modeling are specific to the experimental environment; however, as long as a mathematical representation of flight dynamics is available and a homogeneous flight task is defined, the modeling should be feasible.

The logical follow-up to this research project is its application in a real test situation. Procedures described in this paper are applicable to comparative evaluation of systems based on performance measures of actual pilot response to the systems. Objective evaluation of systems based on pilot response will let designers know how pilot response differs among systems and also the basis of pilot preference in terms of objective performance measures of pilot response.

# REFERENCES

Baldwin, L. C. and Gantz, D. T. (1983). *Pilot Vehicle-Display: Optimal Control Theory Applications in Determination of Relations between Pilot Opinion and Test Measurements*. Maritime Associates, DAAK11-82-C-0115, November 1983.

Baldwin, L. C. and Gantz, D. T. (1984). *Analytical Evaluation of Aircraft / Flight Display System*. Maritime Associates, DAAK11-82-C-0115, August 1984.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, California.

Jex, H. R., McDonnell, J. D., and Phatak, A. V. (1966). *A "Critical" Tracking Task for Man-Machine Research Related to the Operator's Effective Delay Time, Part I: Theory and Experiments with a First-Order Divergent Element*. NASA CR-616, November 1966.

McDonnell, J. D. (1968). *Pilot Rating Techniques for the Estimation and Evaluation of Handling Qualities*. Systems Technology Inc., AFFDL-TR-68-76, December 1968.

SAS (1985). *SAS User's Guide: Statistics, Version 5 Edition*. SAS Institute Inc., Cary, North Carolina.

# APPENDIX A
# PILOT FLIGHT EXPERIENCE QUESTIONNAIRE

Each pilot participating in the experiment completed a personal flight experience questionnaire:

Flight Experience Questionnaire

1. Military flight rating
2. Total flight hours as pilot or copilot
3. Flight hours by aircraft class:
   - Fighter/Attack (VF/VA)
   - Patrol/Transport (VP/VC)
   - Helicopter (Helo)
   - Light Aircraft (Light)
4. Instrument Flight (Actual Hours)
5. Simulator Training Hours
6. Familiarity and Use of Rating Scales
   - Cooper-Harper (C-H)
   - Other
7. Use of Rating Scales in Flight System Evaluation
   - Fixed Base Simulators
   - Moving Base Simulators
   - Aircraft (Class)

The military flight rating for all pilots is Naval Aviator. The following table is compiled from the completed questionnaires:

**Table A-1.** Experience Summary

| Pilot | Total Hours | Flight Hours by Aircraft Class | | | | Inst. Flight Hours | Simul. Training Hours | Rating Scales | |
|---|---|---|---|---|---|---|---|---|---|
| | | VF/ VA | VP/ VC | Helo | Light | | | Type | Use |
| 1 | 2000 | 1800 | 0 | 0 | 200 | 100 | 450 | C-H | VA |
| 2 | 2100 | 2000 | 20 | 20 | 100 | 250 | 0 | C-H | VF<br>VA |
| 3 | 2050 | 10 | 10 | 1900 | 100 | 300 | 80 | C-H | Helo |
| 4 | 2900 | 10 | 2700 | 50 | 150 | 500 | 100 | C-H | VC |
| 5 | 5500 | 0 | 0 | 5400 | 100 | 900 | 300 | C-H | Helo |
| 6 | 1400 | 1400 | 0 | 0 | 25 | 50 | 250 | C-H | VA |
| 7 | 2500 | 2300 | 10 | 10 | 150 | 300 | 200 | C-H | VA |
| 8 | 2300 | 20 | 30 | 2100 | 130 | 350 | 100 | C-H | Helo |
| 9 | 2200 | 2000 | 20 | 15 | 40 | 200 | 50 | C-H | VF |
| 10 | 3000 | 0 | 900 | 1800 | 300 | 1000 | 100 | C-H | Helo |
| 11 | 4000 | 3000 | 600 | 0 | 400 | 550 | 200 | C-H | VA<br>VP |
| 12 | 1610 | 1300 | 10 | 20 | 200 | 300 | 90 | C-H | VF<br>VA |
| 13 | 1400 | 1100 | 0 | 50 | 50 | 200 | 150 | Other | Other |
| 14 | 3000 | 400 | 1800 | 0 | 800 | 300 | Unk | C-H | VA |
| 15 | 1900 | 150 | 1600 | 0 | 100 | 800 | 300 | C-H | VP |

# APPENDIX B
# RUN SEQUENCES EXECUTED BY PILOTS

The number of runs flown by each pilot was subject to time available, and hence varied from pilot to pilot. The number of runs flown by each pilot ranged from 10 to 17. Each run is determined by a particular configuration of system, gain, screen vertical display scale, secondary loading task, and bandwidth of the pitch tracking target. The first two runs for each pilot were used as training runs. These were always based on the K/s system with gain .586, normal screen vertical display scale, and low bandwidth, run both with and without the secondary loading task. The remaining runs "flown" by each pilot are listed below. The configuration variables for a run take on the following values:

System - $K/s$<br>
$K/s(s+1)$<br>
$K/s(s+2)$<br>
$K/s(s+4)$<br>
$K/[s^2 + 2(0.7)(7.8)s + 7.8^2]$ (Poly1)<br>
$K/[s^2 + 2(0.7)(16)s + 16^2]$ (Poly2)<br>
$K/s^2$

Gain - See Table 2

Lambda (Secondary Loading Task) - On or Off

Bandwidth (of Pitch Tracking Target) - Low or High

Screen (Vertical Display Scale) - Normal or Wide

C-H (Cooper-Harper Rating) - Pilot's Subjective Rating (See Appendix C)

Time - Run Duration in Seconds

Notes on Data:

1. The data for Pilot 8 were collected in two sessions; the data for all other pilots were collected in a single session.

2. The Cooper-Harper rating for Run 4, Pilot 11 failed to be recorded and is marked as missing. Some pilots chose to give non-integer values for the Cooper-Harper rating.

3. The runs marked "Failed Run" were terminated abnormally either due to problems with the IRIS 3000 computer or due to the pilot losing control of the tracking task. The large number of failed runs for Pilot 10 is due to a computer hardware problem.

## Pilot #1

| Run | System | Gain | Lambda | Band | Screen | C-H | Time |
|---|---|---|---|---|---|---|---|
| 1 | K/s | .586 | Off | Low | Normal | 2.5 | 113 |
| 2 | K/s | .586 | On | Low | Normal | 5.0 | 163 |
| 3 | K/s(s+4) | 58.6 | Off | Low | Normal | 3.0 | 135 |
| 4 | K/s(s+4) | 58.6 | On | Low | Normal | 4.5 | 204 |
| 5 | K/s | .293 | Off | Low | Normal | 4.5 | 134 |
| 6 | K/s | .293 | On | Low | Normal | 6.0 | 194 |
| 7 | Poly1 | 21.5 | Off | Low | Normal | 5.0 | 197 |
| 8 | Poly1 | 21.5 | On | Low | Normal | 6.0 | 204 |
| 9 | K/s(s+2) | 35.2 | Off | Low | Normal | 4.0 | 154 |
| 10 | Failed Run | | | | | | |
| 11 | K/s | 5.86 | Off | Low | Normal | 3.0 | 200 |
| 12 | Failed Run | | | | | | |

## Pilot #2

| Run | System | Gain | Lambda | Band | Screen | C-H | Time |
|---|---|---|---|---|---|---|---|
| 1 | K/s | .586 | Off | Low | Normal | 4.0 | 118 |
| 2 | K/s | .586 | On | Low | Normal | 5.0 | 87 |
| 3 | Failed Run | | | | | | |
| 4 | K/s(s+4) | 117 | On | Low | Normal | 7.0 | 131 |
| 5 | K/s(s+4) | 83.8 | On | Low | Normal | 5.0 | 123 |
| 6 | Poly1 | 35.2 | Off | Low | Normal | 7.0 | 124 |
| 7 | Poly1 | 35.2 | On | Low | Normal | 7.0 | 115 |
| 8 | $K/s^2$ | .586 | Off | Low | Normal | 8.0 | 117 |
| 9 | $K/s^2$ | .293 | On | Low | Normal | 8.0 | 108 |
| 10 | K/s(s+1) | 35.2 | Off | Low | Normal | 4.0 | 85 |
| 11 | K/s(s+1) | 35.2 | On | Low | Normal | 4.0 | 112 |
| 12 | K/s | 5.86 | Off | Low | Normal | 5.0 | 113 |
| 13 | K/s | 5.86 | On | Low | Normal | 6.0 | 137 |
| 14 | Poly1 | 83.8 | Off | Low | Normal | 7.0 | 119 |
| 15 | Poly1 | 83.8 | On | Low | Normal | 7.0 | 134 |
| 16 | K/s(s+1) | 35.2 | Off | Low | Normal | 4.0 | 102 |
| 17 | K/s(s+1) | 35.2 | On | Low | Normal | 4.0 | 134 |

## Pilot #3

| Run | System | Gain | Lambda | Band | Screen | C-H | Time |
|---|---|---|---|---|---|---|---|
| 1 | K/s | .586 | Off | Low | Normal | 4.0 | 94 |
| 2 | K/s | .586 | On | Low | Normal | 5.5 | 151 |
| 3 | K/s(s+4) | 58.6 | Off | Low | Normal | 3.0 | 147 |
| 4 | K/s(s+4) | 58.6 | On | Low | Normal | 4.3 | 170 |
| 5 | K/s(s+2) | 21.5 | Off | Low | Normal | 2.8 | 102 |
| 6 | K/s(s+2) | 17.6 | On | Low | Normal | 4.0 | 133 |
| 7 | Poly2 | 35.2 | On | Low | Normal | 8.0 | 164 |
| 8 | Poly1 | 35.2 | On | Low | Normal | 8.5 | 132 |
| 9 | Poly1 | 17.6 | On | Low | Normal | 5.0 | 123 |
| 10 | K/s(s+2) | 8.38 | On | Low | Normal | 7.0 | 152 |
| 11 | Poly2 | 215 | On | Low | Normal | 8.5 | 174 |
| 12 | Poly2 | 35.2 | On | Low | Normal | 8.0 | 177 |

## Pilot #4

| Run | System | Gain | Lambda | Band | Screen | C-H | Time |
|---|---|---|---|---|---|---|---|
| 1 | K/s | .586 | Off | Low | Normal | 4.0 | 160 |
| 2 | K/s | .586 | On | Low | Normal | 4.0 | 152 |
| 3 | K/s(s+4) | 58.6 | Off | Low | Normal | 3.5 | 263 |
| 4 | K/s(s+4) | 58.6 | On | Low | Normal | 3.5 | 216 |
| 5 | K/s | 8.38 | On | Low | Normal | 8.0 | 107 |
| 6 | Poly1 | 35.2 | Off | Low | Normal | 7.0 | 197 |
| 7 | Poly1 | 83.8 | On | Low | Normal | 8.0 | 88 |
| 8 | K/s(s+1) | 83.8 | On | Low | Normal | 7.0 | 134 |
| 9 | Failed Run | | | | | | |
| 10 | $K/s^2$ | .293 | On | Low | Normal | 10.0 | 169 |
| 11 | Poly1 | 17.6 | Off | Low | Normal | 6.0 | 162 |
| 12 | Poly1 | 17.6 | On | Low | Wide | 7.0 | 217 |

## Pilot #5

| Run | System | Gain | Lambda | Band | Screen | C-H | Time |
|---|---|---|---|---|---|---|---|
| 1 | K/s | .586 | Off | Low | Normal | 3.0 | 145 |
| 2 | K/s | .586 | On | Low | Normal | 5.0 | 157 |
| 3 | K/s(s+4) | 58.6 | Off | Low | Normal | 2.0 | 237 |
| 4 | K/s(s+4) | 58.6 | On | Low | Normal | 2.5 | 226 |
| 5 | K/s | .586 | On | High | Normal | 4.0 | 175 |
| 6 | $K/s^2$ | .586 | On | Low | Wide | 8.5 | 141 |
| 7 | K/s | .586 | On | High | Wide | 5.0 | 164 |
| 8 | Poly1 | 21.5 | Off | Low | Normal | 3.0 | 168 |
| 9 | Failed Run | | | | | | |
| 10 | Poly1 | 21.5 | On | Low | Normal | 4.0 | 190 |
| 11 | K/s(s+4) | 58.6 | On | High | Normal | 3.0 | 157 |
| 12 | K/s(s+4) | 58.6 | On | High | Wide | 5.0 | 181 |
| 13 | K/s(s+4) | 117 | On | Low | Wide | 5.0 | 165 |
| 14 | $K/s^2$ | .586 | On | High | Normal | 9.0 | 131 |

## Pilot #6

| Run | System | Gain | Lambda | Band | Screen | C-H | Time |
|---|---|---|---|---|---|---|---|
| 1 | K/s | .586 | Off | Low | Normal | 3.0 | 136 |
| 2 | K/s | .586 | On | Low | Normal | 4.0 | 190 |
| 3 | K/s(s+4) | 58.6 | On | Low | Normal | 4.0 | 232 |
| 4 | K/s(s+4) | 58.6 | On | Low | Wide | 4.5 | 161 |
| 5 | Poly1 | 21.5 | On | Low | Wide | 5.0 | 160 |
| 6 | $K/s^2$ | 586 | On | High | Wide | 8.0 | 149 |
| 7 | $K/s^2$ | 586 | On | Low | Normal | 7.0 | 221 |
| 8 | Poly1 | 21 5 | On | High | Wide | 7.0 | 183 |
| 9 | Poly1 | 21 5 | On | High | Normal | 5.0 | 166 |
| 10 | K/s(s+4) | 58 6 | On | High | Normal | 3.0 | 156 |

Pilot #7

| Run | System | Gain | Lambda | Band | Screen | C-H | Time |
|---|---|---|---|---|---|---|---|
| 1 | K/s | .586 | Off | Low | Normal | 3.0 | 131 |
| 2 | K/s | .586 | On | Low | Normal | 4.0 | 162 |
| 3 | K/s | .293 | Off | Low | Normal | 3.0 | 141 |
| 4 | K/s | .293 | On | Low | Normal | 4.5 | 154 |
| 5 | Poly2 | 35.2 | Off | Low | Normal | 3.0 | 220 |
| 6 | Poly2 | 35.2 | On | Low | Normal | 5.0 | 201 |
| 7 | K/s(s+2) | 35.2 | Off | Low | Normal | 3.0 | 199 |
| 8 | K/s(s+2) | 35.2 | On | Low | Normal | 5.0 | 208 |
| 9 | K/s(s+2) | 21.5 | On | Low | Normal | 4.5 | 233 |
| 10 | K/s(s+2) | 21.5 | Off | Low | Normal | 2.5 | 146 |

Pilot #8

| Run | System | Gain | Lambda | Band | Screen | C-H | Time |
|---|---|---|---|---|---|---|---|
| 1 | K/s | .586 | Off | Low | Normal | 3.0 | 209 |
| 2 | K/s | .586 | On | Low | Normal | 5.0 | 188 |
| 3 | K/s(s+1) | 35.2 | Off | Low | Normal | 3.0 | 185 |
| 4 | K/s(s+1) | 35.2 | On | Low | Normal | 4.0 | 274 |
| 5 | K/s | .586 | Off | Low | Normal | 5.0 | 118 |
| 6 | K/s | .586 | On | Low | Normal | 6.0 | 186 |
| 7 | K/s(s+4) | 35.2 | On | Low | Normal | 5.0 | 194 |
| 8 | K/s(s+4) | 17.6 | On | Low | Normal | 6.0 | 190 |
| 9 | K/s(s+4) | 58.6 | On | Low | Normal | 4.0 | 185 |
| 10 | $K/s^2$ | 586 | On | Low | Normal | 10.0 | 212 |
| 11 | $K/s^2$ | 1 17 | On | Low | Normal | 8.0 | 164 |
| 12 | $K/s^2$ | 293 | On | Low | Normal | 9.0 | 189 |
| 13 | K/s | 1 17 | On | Low | Normal | 4.0 | 240 |
| 14 | Poly1 | 17.6 | On | Low | Normal | 5.0 | 206 |

Pilot #9

| Run | System | Gain | Lambda | Band | Screen | C-H | Time |
|---|---|---|---|---|---|---|---|
| 1 | K/s | .586 | Off | Low | Normal | 4.0 | 153 |
| 2 | K/s | .586 | On | Low | Normal | 5.0 | 219 |
| 3 | K/s(s+4) | 35.2 | Off | Low | Normal | 3.0 | 210 |
| 4 | K/s(s+4) | 35.2 | On | Low | Normal | 4.0 | 204 |
| 5 | Poly2 | 35.2 | On | Low | Normal | 3.0 | 188 |
| 6 | $K/s^2$ | .586 | On | Low | Normal | 8.0 | 191 |
| 7 | K/s(s+4) | 17.6 | On | Low | Normal | 3.0 | 193 |
| 8 | K/s | .293 | On | Low | Normal | 4.5 | 192 |
| 9 | Poly1 | 35.2 | On | Low | Normal | 3.0 | 178 |
| 10 | Poly2 | 35.2 | On | Low | Normal | 3.0 | 175 |
| 11 | Poly1 | 17.6 | On | Low | Normal | 4.0 | 195 |

Pilot #10

| Run | System | Gain | Lambda | Band | Screen | C-H | Time |
|---|---|---|---|---|---|---|---|
| 1 | K/s | .586 | Off | Low | Normal | 4.0 | 207 |
| 2 | K/s | .586 | On | Low | Normal | 6.0 | 234 |
| 3 | Failed Run | | | | | | |
| 4 | Poly2 | 83.8 | On | Low | Wide | 6.0 | 208 |
| 5 | Failed Run | | | | | | |
| 6 | K/s | .586 | On | High | Normal | 5.0 | 202 |
| 7 | Poly1 | 35.2 | On | High | Normal | 6.0 | 212 |
| 8 | Failed Run | | | | | | |
| 9 | Poly1 | 35.2 | On | Low | Normal | 5.0 | 190 |
| 10 | Failed Run | | | | | | |
| 11 | K/s(s+4) | 35.2 | On | Low | Wide | 5.0 | 188 |
| 12 | $K/s^2$ | .586 | On | High | Normal | 9.0 | 186 |
| 13 | K/s(s+2) | 35.2 | On | Low | Wide | 5.0 | 172 |

Pilot #11

| Run | System | Gain | Lambda | Band | Screen | C-H | Time |
|---|---|---|---|---|---|---|---|
| 1 | K/s | .586 | Off | Low | Normal | 3.0 | 174 |
| 2 | K/s | .586 | On | Low | Normal | 6.0 | 245 |
| 3 | K/s(s+4) | 35.2 | On | Low | Normal | 2.5 | 224 |
| 4 | Poly1 | 17.6 | On | High | Wide | Missing | 282 |
| 5 | K/s(s+1) | 35.2 | On | High | Normal | 7.0 | 180 |
| 6 | K/s(s+1) | 35.2 | On | Low | Normal | 3.0 | 228 |
| 7 | K/s(s+2) | 35.2 | On | Low | Normal | 2.0 | 186 |
| 8 | K/s(s+1) | 35.2 | On | High | Wide | 7.0 | 246 |
| 9 | K/s(s+2) | 35.2 | On | High | Normal | 7.0 | 195 |
| 10 | Poly1 | 58.6 | On | Low | Normal | 8.0 | 288 |

Pilot #12

| Run | System | Gain | Lambda | Band | Screen | C-H | Time |
|---|---|---|---|---|---|---|---|
| 1 | K/s | .586 | Off | Low | Normal | 5.0 | 142 |
| 2 | K/s | .586 | On | Low | Normal | 6.0 | 170 |
| 3 | K/s(s+4) | 35.2 | On | Low | Normal | 5.0 | 154 |
| 4 | K/s(s+4) | 35.2 | Off | Low | Normal | 4.0 | 101 |
| 5 | $K/s^2$ | .293 | On | Low | Normal | 10.0 | 75 |
| 6 | K/s | .293 | On | Low | Normal | 6.0 | 160 |
| 7 | K/s(s+2) | 58.6 | On | Low | Normal | 5.0 | 182 |
| 8 | K/s(s+1) | 58.6 | On | Low | Normal | 8.0 | 193 |
| 9 | K/s(s+2) | 35.2 | On | Low | Normal | 4.5 | 156 |
| 10 | Failed Run | | | | | | |
| 11 | K/s(s+2) | 35.2 | Off | Low | Normal | 4.5 | 164 |

Pilot #13

| Run | System | Gain | Lambda | Band | Screen | C-H | Time |
|---|---|---|---|---|---|---|---|
| 1 | K/s | .586 | Off | Low | Normal | 3.0 | 197 |
| 2 | K/s | .586 | On | Low | Normal | 5.0 | 199 |
| 3 | K/s(s+4) | 35.2 | Off | Low | Normal | 3.0 | 175 |
| 4 | K/s(s+4) | 35.2 | On | Low | Normal | 4.0 | 174 |
| 5 | Poly2 | 35.2 | On | Low | Normal | 7.0 | 189 |
| 6 | $K/s^2$ | .586 | On | Low | Normal | 9.0 | 177 |
| 7 | K/s(s+1) | 17.6 | On | Low | Normal | 3.0 | 183 |
| 8 | K/s(s+4) | 58.6 | On | Low | Normal | 4.0 | 187 |
| 9 | K/s(s+2) | 17.6 | On | Low | Normal | 4.0 | 193 |
| 10 | K/s(s+1) | 35.2 | On | Low | Normal | 5.0 | 174 |
| 11 | Failed Run | | | | | | |
| 12 | Poly1 | 35.2 | On | Low | Normal | 6.0 | 182 |

Pilot #14

| Run | System | Gain | Lambda | Band | Screen | C-H | Time |
|---|---|---|---|---|---|---|---|
| 1 | K/s | .586 | Off | Low | Normal | 4.0 | 172 |
| 2 | K/s | .586 | On | Low | Normal | 6.0 | 198 |
| 3 | K/s(s+4) | 35.2 | On | Low | Normal | 4.0 | 196 |
| 4 | $K/s^2$ | .586 | On | High | Wide | 9.0 | 160 |
| 5 | K/s(s+1) | 35.2 | On | High | Normal | 5.0 | 185 |
| 6 | K/s | .586 | On | Low | Wide | 5.0 | 186 |
| 7 | Poly1 | 35.2 | On | Low | Wide | 7.0 | 183 |
| 8 | K/s(s+2) | 35.2 | On | Low | Normal | 4.0 | 128 |
| 9 | Poly2 | 35.2 | On | High | Normal | 6.0 | 193 |
| 10 | Poly2 | 35.2 | On | Low | Normal | 3.0 | 198 |

Pilot #15

| Run | System | Gain | Lambda | Band | Screen | C-H | Time |
|---|---|---|---|---|---|---|---|
| 1 | K/s | .586 | Off | Low | Normal | 5.0 | 168 |
| 2 | K/s | .586 | On | Low | Normal | 7.0 | 180 |
| 3 | K/s(s+4) | 35.2 | On | Low | Normal | 6.0 | 173 |
| 4 | Poly1 | 17.6 | On | High | Wide | 7.0 | 190 |
| 5 | K/s(s+1) | 35.2 | On | Low | Normal | 5.0 | 200 |
| 6 | K/s(s+1) | 35.2 | On | High | Normal | 6.0 | 189 |
| 7 | K/s(s+2) | 35.2 | On | Low | Normal | 4.0 | 208 |
| 8 | K/s(s+1) | 35.2 | On | High | Wide | 8.0 | 241 |
| 9 | K/s(s+2) | 35.2 | On | High | Wide | 6.0 | 252 |
| 10 | K/s(s+4) | 35.2 | On | High | Normal | 4.0 | 219 |
| 11 | $K/s^2$ | .586 | On | Low | Normal | 10.0 | 157 |

# APPENDIX C
# SUBJECTIVE EVALUATION PROCEDURE USING COOPER-HARPER RATING

Each pilot was briefed on the objectives and nature of the experiment including the characteristics of each of the systems planned for the session. The Cooper-Harper rating scale was reviewed, and a set of questions relating to the evaluation were discussed. All pilots except one had been trained in the use of the Cooper-Harper rating scale and had used it in various aircraft projects.

The tracking task was described as follows:

> The primary task is to track the target as accurately as possible, keeping the center of the target within the "diamond" symbol. It is not essential that your "wings" are level. The secondary objective is to keep the "wings" as level as possible.

Each pilot was asked to evaluate the system flown in terms of the following elements:

Response Characteristics - System time constants, overshoot or undershoot, and lag.

Ease and Precision of Control - System damping, stick (controller) acceptability, target motion.

Demands on Pilot - Complexity and difficulty of tracking tasks.
Effect of secondary task on primary task.

Effect of System "Deficiencies" on Performance - Overshoot, control response, secondary "roll" control.

Each pilot was asked to also provide an overall rating based on the Cooper-Harper rating scale adapted to this experiment:

**Table C-1.** Adapted Cooper-Harper Rating Scale

| System Characteristics | Pilot Compensation For Performance of Task | Cooper-Harper Rating |
|---|---|---|
| Excellent | Not a Factor | 1 |
| Good | Not a Factor | 2 |
| Fair | Minimal | 3 |
| Minor Defects | Moderate | 4 |
| Defects Objectionable | Considerable | 5 |
| Very Objectionable | Extensive | 6 |
| Major Performance Problem | Barely Perform Task | 7 |
| Major Control Problem | Difficult to Control | 8 |
| Major Deficiency | Barely Controllable | 9 |
| Major Deficiency | Uncontrollable | 10 |

# APPENDIX D
# STATISTICAL DISCRIMINATION ALGORITHMS

## D-1. SAS Procedure STEPDISC

STEPDISC, Stepwise Discrimination, performs stepwise classification variable selection useful as a precursor to estimating good discrimination rules for a grouping variable. The procedure measures the classification power of a variable via the significance level of an F-test for equality of group means from an Analysis of Covariance where the classification variables already chosen are used as covariates and the current classification variable under consideration is used as the dependent variable.

At each stage, the variable with the most significant F-test for equality of group means is added to the list of variables in the model; then, each selected variable is tested for removal in light of the most recently added variable. The algorithm continues until no classification variables can be added and none removed based on user-defined significance levels for entry and removal (or alternatively, based on user-defined levels for the squared partial correlation coefficient for predicting the classification variable under consideration from the grouping variable controlling for the effects of the already selected variables).

The significance level of the F-test is based on assuming a multivariate Gaussian distribution for the classification variables within each group with a common covariance matrix for each group. The SAS procedure can also be used to do strictly forward selection or backward selection.

## D-2. SAS Procedure CANDISC

CANDISC, Canonical Discrimination Analysis, is related to Principal Components and Canonical Correlations. Given a grouping variable and some quantitative classification variables, CANDISC derives linear combinations of the quantitative classification variables which are highly correlated with the group variable. These linear combinations summarize the between-class variation in much the same way that Principal Components Analysis summarizes total variation; in fact, Canonical Discrimination Analysis is equivalent to performing Principal Component Analysis on the class means of standardized variables. Canonical Discrimination Analysis is also equivalent to Canonical Correlation Analysis between the quantitative classification variables and a set of dummy variables encoding the group variable.

The linear combination with the highest possible multiple correlation with the grouping variable is called the first canonical variable; the coefficients in the linear combination are called the canonical coefficients or weights; the multiple correlation is called the canonical correlation. The second canonical variable is obtained by finding the linear combination uncorrelated with the first canonical variable that has the highest possible multiple correlation with the grouping variable and so on for the third, fourth, etc. The maximum number of canonical variables is the number of classification variables or the number of groups minus one, whichever is smaller.

The inference in this procedure is based on assuming a multivariate Gaussian distribution for the quantitative classification variables within each group with a common covariance matrix for each group.

### D-3. CART (Classification and Regression Trees)

CART computes a discrimination rule which can be represented in a binary hierarchical tree structure. The rule consists of a sequence of binary questions (yes/no) based on the value of a single classification variable. Each of these questions results in splitting the observations into two subsets each of which are split into two smaller subsets and so on. Since the questions are binary, these repeated splits into subsets can be represented in a tree structure where each subset becomes a node of the tree. Each node is then associated with a single variable and has two descendant nodes; the value of this variable for a given observation determines whether that observation is further classified into the left or right descendant node. Each binary question is chosen to maximize the homogeneity of the observations in each descendent node in terms of group membership. The "bottom" nodes of the tree, called terminal nodes, are each assigned a group. Then, this tree structure can serve as a discrimination rule as follows. At each node of the tree, an observation goes either to the right or the left descendant node depending upon the value of the variable associated with the parent node. The observation continues down the tree by checking the value of the appropriate classification variable at each node. Eventually, the observation lands in a terminal node and is classified into the group associated with the terminal node.

The construction of such a classification tree involves three elements:

1. How to select each binary question to produce each split.
2. How to decide whether a node is terminal or not.
3. How to assign a group to each terminal node.

The selection of the binary questions is based upon an impurity function I. This function is defined over all vectors $(p_1,p_2,\ldots,p_J)$, where J is the number of groups, $p_j \geq 0$ for each j, and $\Sigma_j\, p_j = 1$. It takes on its maximum value at (1/J,1/J,...,1/J) and its minimum at points (1,0,...,0), (0,1,...,0), . . . , (0,0,...,1). Also, it is a symmetric function of $(p_1,p_2,...,p_J)$. In terms of the impurity function, an impurity index can be defined for each node:

$$\text{Impurity Index of Node } t = I(t) = I(p(1\,|\,t),p(2\,|\,t),...,p(J\,|\,t))$$

where $p(j\,|\,t)$ is the proportion of group j observations in node t. Thus, the impurity index is a function of the proportion of each group in a node. The properties of the impurity function make the impurity index of a node largest when all groups are equally mixed in a node and smallest when the subset contains a single group. If a binary question sends a proportion $p_R$ of the data to the right node $t_R$ and a proportion $p_L$ to the left node $t_L$, then the decrease in impurity becomes a measure of the goodness of a split:

$$\text{Goodness of Split} = I(t) - p_R I(t_R) - p_L I(t_L).$$

At each node, CART searches through all classification variables one by one choosing that split which maximizes the goodness of split over all possible splits. For a quantitative variable with N distinct values, there are at most N distinct splits generated by $x \le c$ where c takes values halfway between consecutive distinct values of the classification variable x. For a qualitative variable with L distinct values, there are $2^{L-1}$ possible splits corresponding to the number of distinct subsets. Having found the best split for each variable, it compares the best splits to find the best split over all. Studies have shown that within a wide range of splitting criteria, the properties of the final tree is insensitive to the choice of impurity function.

The type of splits can be generalized in several ways. With quantitative variables, one can consider linear combinations of the classification variables — whether or not they are smaller or greater than a given value. With qualitative variables, one can consider splits based on Boolean logic statements about the simultaneous occurrence of two or more variables in certain subsets.

The decision of whether or not a node is terminal is based on estimates of the misclassification rate of the tree. The misclassification rate can be viewed as the probability that a "new" observation "run" down the tree is classified into the wrong group. There are several ways to estimate this misclassification rate based on the observed data.

One estimator is called the resubstitution estimate. Here, you estimate the misclassification rate using the same data used to construct the tree. The percentage of misclassified observations in your original set is the estimate of the misclassification rate. The problem with using this estimator is that all discrimination algorithms, either directly or indirectly, construct their discrimination rules via minimizing this measure of misclassification. Thus, the resubstitution estimate is biased downward as an estimate of the misclassification rate over the population.

Another estimator is called the test sample estimate. Here, before constructing your tree, you randomly divide your data into two parts — one part is used to construct the tree and the other is used to estimate the misclassification rate by the percentage of misclassified observations. The first set is called the learning sample and the second called the test sample. The problem with this estimator is that it reduces your effective sample size, which is no problem with large data sets but can be a problem with smaller ones. The final estimator is called the V-fold cross-validation estimate. Here, the data set is randomly divided into V subsets of about the same size. For each subset, a tree is grown using all observations not in the subset and then the percentage of misclassified observations in the subset is computed. The average misclassification rate over all subsets is then the cross-validation estimate.

CART uses either the test sample estimate or the cross-validation estimate to determine when a node is terminal or equivalently to select the right size tree. First, CART grows a large tree — larger than the data actually warrants as far as information content. CART continues to compute splits until all terminal nodes contain less than a user-specified number of observations or until all observations in the terminal nodes are from the same group. Then, CART selectively recombines nodes upward in this large original tree to create a nested sequence of trees with smaller and smaller number of terminal nodes. CART then estimates the misclassification rate of each tree in the sequence using either a test sample or V-fold cross-validation estimate. The final tree selected (and thus the terminal nodes selected) is the tree in the sequence with the smallest estimated misclassification rate. (The resubstitution estimate cannot be used because it decreases as the number of nodes increases.)

Given the set of terminal nodes, a group assignment rule assigns a group to every terminal node. One rule that can be used is to assign the group with the largest membership in the node. This rule minimizes the misclassification rate within the node. One can also assign a cost or loss to misclassifying a group j observation as a group i observation. Then, one can choose the group for each terminal node that minimizes the expected misclassification cost.

The tree structured approach to discrimination analysis as implemented in CART has several advantages over other techniques:

1. It can handle any type of classification variable — either qualitative or quantitative.

2. The final discrimination rule has a simple form — a sequence of yes/no questions. It is easy to understand and to use in determining the predictive structure of the data.

3. It takes advantage of conditional information in handling nonhomogeneous relationships. This allows for finding different rules to differentiate between different subsets of groups.

4. It automatically does stepwise variable selection.

5. It automatically provides a "good" estimate of misclassification rate.

6. It is invariant under all monotone transformations of the individual classification variables.

7. It is extremely robust against outliers and misclassified observations. It is not based on any distributional assumptions for the classification variables within each group.

For further details on CART, see Breiman, Friedman, Olshen, and Stone (1984).

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

## REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
Final Report

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
Evaluation and Estimation of Handling Qualities via Statistical Modeling of Pilot Response Data

5. TYPE OF REPORT & PERIOD COVERED
15 July 1989-28 February 1991

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Donald T. Gantz
Lawrence C. Baldwin
Linda J. Davis

8. CONTRACT OR GRANT NUMBER(s)
N00014-89-J-3146

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Center for Computational Statistics, George Mason University
Fairfax, VA 22030

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS
Office of Naval Research
800 N. Quincy Street
Arlington, VA 22217-5000

12. REPORT DATE
November, 1991

13. NUMBER OF PAGES
105

14. MONITORING AGENCY NAME & ADDRESS (If different from Controlling Office)
Naval Test Pilot School
Naval Air Test Center
Patuxent River, Maryland

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
Cooper-Harper, Flight Display Evaluation, Flight Simulation, Flight Test, Handling Qualities, Pilot Modeling, Pilot Performance, Pilot Response

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This report describes a research project which measured pilot response to seven control systems simulating different handling qualities, quantitatively evaluated and compared the systems based on these measurements, and compared the quantitative system evaluation based on measured pilot performance with a qualitative evaluation using the Cooper-Harper technique.

Pilot performance is determined through analysis of objective dynamic measurements of pilot response typical of flight test environments. In short, the methodology specifies a general approach for condensing the typically huge mound of measured test data accumulated during flight simulation experiments into meaningful quantities for system evaluation. The key element in the methodology is statistical

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20. (continued) modeling of a law for pilot control. Statistical modeling of pilot control provides an assessment of pilot performance in terms of standard statistical estimation parameters. The methodology requires that this control model be used to compute control input in a closed loop tracking task; the accuracy of the control model in perfoming this task is an important measure of pilot performance relevant to system evaluation. In addition, these paramenters computed from the dynamic measurements of pilot performance are shown to enhance understanding of the aspects of the handling qualities underlying subjective rating techniques such as Cooper-Harper.